IJCSIS Vol. 10 No. 5, May 2012
ISSN 1947-5500

International Journal of Computer Science & Information Security

Editorial
Message from Managing Editor

International Journal of Computer Science and Information Security (IJCSIS) publishes original research works and reviewed articles in all areas of computer science including emerging topics like cloud computing, software development etc. The journal promotes insight and understanding of the state of the art and trends in computing technology and applications.

IJCSIS solicits authors/researchers/scholars to contribute to the journal by submitting articles that illustrate research results, projects, surveying works and industrial experiences. IJCSIS helps academia to rapidly and continuously publish academic work to sustain or further one's career.

For complete details about IJCSIS archives publications, abstracting/indexing, editorial board and other important information, please refer to IJCSIS homepage. IJCSIS appreciates all the insights and advice from authors/readers and reviewers. Indexed by the following International Agencies and institutions: EI, Scopus, DBLP, DOI, ProQuest. Average acceptance for the period January-May 2012 is 26%.

We look forward to receive your valuable papers. The topics covered by this journal are diverse. (See monthly Call for Papers). If you have further questions please do not hesitate to contact us at ijcsiseditor@gmail.com. Our team is committed to provide a quick and supportive service throughout the publication process.

A complete list of journals can be found at:

http://sites.google.com/site/ijcsis/

IJCSIS Vol. 10, No. 5, May 2012 Edition

ISSN 1947-5500 © IJCSIS, USA & UK.

Journal Indexed by (among others):

IJCSIS EDITORIAL BOARD

TABLE OF CONTENTS

Multidimensional analysis in intelligence business systems

Elma Zanaj , Ledion Liço, Indrit Enesi
Electronic and Telecommunication Department
Polytechnic University of Tirana
Tirana, Albania

***Abstract*— The purpose of this study is to create an OLTP (Online Transaction Processing) and a DW (Data Warehouse) in order to make it simpler the extraction of various reports and to take information from different systems. Another purpose is to make a comparison between different OLAP (Online Analytical Processing)**
technologies for a large number of records. We will compare HOLAP and ROLAP technologies and their performance will be evaluated. For this reason it will be tested on DW a query by using ROLAP and to the intelligent cubes that will be created by using HOLAP for a considerable number of records and the system response time will be analyzed.

Keywords::information, systems, business, analysis

I. INTRODUCTION

This study will compare the systems OLTP and DW [1] systems for analyzing data and creating different reports for users. [2]. Currently it is difficult to generate reports due to the complexity of systems and because of the existence of many systems in the corporations. There are several reports in Gestcomm application but they do not offer the opportunity to add new reports. So in such cases is needed to write long SQL codes that require long time and have a high complexity. Our goal is to simplify and to increase the quality of information extracted from these systems. During this study we will compare the OTLP and DW systems for analyzing the data and creating various reports, for users and for a large number of records.

We will create a DW for a company that uses an OTLP system and will analyze the advantages that this system creates in providing BI (Business Intelligence) to users. For this reason will be tested a query in the created DW and to the OLAP cubes. We will analyze the systems response time for a considerable number of records. We will make the review of a company whose main activity is the importing and the selling of various goods.

The company operates in the retail sector and trades goods in a shopping center that has different service departments. Goods are classified into categories according to the department that those belong to.

The company uses a transactional information system to perform the transfer operations to another company. This system is based on a relational database based on ORACLE [3]. The application used is Gestcomm and the functions it performs are: a) transfer of goods, b) returning back of goods, c) entry of goods, d) stock management. During the evening there is an information exchange flow between different systems. The company we are reviewing pass the data for the new articles in application database in the Oracle when the Gestcomm application is based on, and from this database the company sales are passed to the other systems of the company. Sales are passed to the Oracle database from Firebird databases where the application of sales is based on, POS2000, and the information for the latest articles and promotions are transferred to the latter by the Oracle database.

After a questionnaire done to managers of the shopping center were needed to do these reports:

•To report accurately on sales based on daily, weekly, monthly and yearly base. This information must be obtained under certain dimensions, categories of goods, clients, set discounts, different sellers, different seasons.

•Report on progress of various promotions located in the shopping center.

•Financial reports for the amount of articles sold in different selling point of the shopping center and the incomings earned.

•Various reports on the clients preferences according to the time and the types of goods favorites.

These reports should be daily and must be taken by each manager and financier. As you will see from our results in case of a query to the transactional system, more tables are connected and the time that we need to get the results is much longer than in the case of DW system.

Another important feature to emphasize is that by increasing the number of rows in the transactional systems the time become longer substantially, while in the case of DW system with the increasing of rows the time become longer but a few. So we have a better performance of this reporting system.

The remainder of the paper is organized as follows. In Section 2, is described the actual system that is used in the company that we are studding ,Section 3 outlines the steps for modeling the data warehouse, while Section 4 shows the evaluation of the performance of the systems by using our simulation results. Finally, the Conclusions concludes the paper.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

II. THE CURRENT SYSTEM IN THE COMPANY A

Company A has as its principal activity the importing and selling of different goods. The company operates in the retail sector. The company has another company partner, B, from which it receives most of the goods and has transactional information system that connects to this company for various transactions.

Company A trades goods in a shopping center divided by service departments. Goods are classified into categories at departments. This system is based on a relational database located in ORACLE.

The application used is Gestcomm and the types of movements that are performed are:

1. Transferring of goods
2. Giving back of goods to the supplier
3. Entering goods
4. Put in order the stock

This application is supplied with new data for the goods from another application DROMOS by the partner company B. This application is supplied with the sale from another application POS2000 by which the sales are done in the shopping center. This application of sales is located in a database FIREBIRD. Through POS2000 is performed even the management of the client card that company A provides.

Each customer receives certain benefits when reaches a certain points and obtains certain promotions such as discounts of 5% or 10% and 15% on the day of his birthday. The GHELDA application serves to place the items that will be on promotion together with the percentage and the name of promotion. During the evening there is the flow of information between different systems as in Fig.1. The company B passes the data of new goods in the application database Oracle when is placed Gestcomm application and from this database the sales are pass to the systems of B. The sales are passed to the Oracle database from the Firebird database where the application of sales POS2000 is placed and the information on the latest articles and promotions of goods is transferred to the latter by the Oracle database.

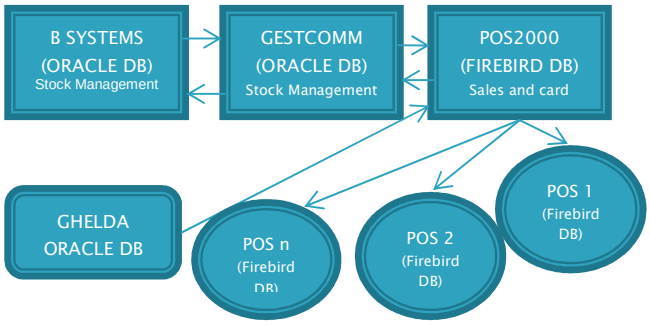


Figure 1. The flow of information of the company A

In the center are located selling points which have a Firebird database which communicates with a central database and pass the information on sales and new customers and receive information on sales and promotions from the latter. Also in the center are located several systems for measuring the flow of visitors and an access control system for sellers and managers from which is taken the information.

The purpose of this study is to create an OLAP system and a data warehouse DW in order to make simple the extraction of various reports and the information from different systems.

Currently there is a difficulty in generating reports due to the complexity of systems and the existence of many systems in this corporation. There are few reports on Gestcomm application but there is no option for adding new reports and this needs to be written SQL code that require long time and have high complexity. The aim would be to simplify and grow the quality of information extracted from these systems.

III. MODELING OF A DATA WAREHOUSE

A. Conceptual Stage

The first step in constructing a DW is the conceptual stages during which will be studied the data that will populate the DW. Therefore is required a detailed analysis of transactional systems and databases to see where is store the information that we will serve us for reporting. Information will be obtained from various sources that in our case are: system sales POS200, stock control system Gestcommand, various documents that hold additional information that managers need for different reports. To achieve the identification of all the necessary information, is done a detailed analysis of operating tables and all fields that will be imported into our DW. Fig.2 shows some of the tables used.
The information is taken from the tables and from the two systems and some documents in excel that the managers save. The tables have primary keys and external keys that connect each table with the others [4].

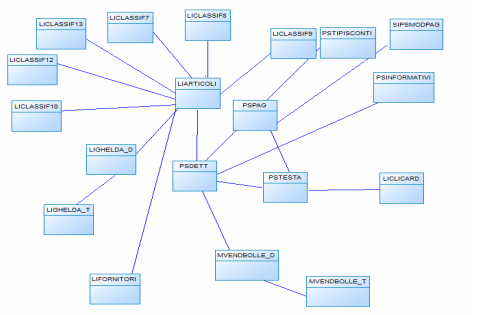


Figure 2. The tables used by the transactional systems

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. The construction phase of the DW

At this stage is constructed the DW by using a star model. In center will be a fact table which will be linked to the tables of dimensions and segments from which it dimension this fact table. At the fact table all the data are not elaborate, they are collected from all operating systems and this is the main source of information. After a detailed analysis of information what we will need to create reports, is builded the DW scheme, Fig.3. For this model is used the Power Designer program which helps in constructing databases and from it can easily export to SQL scripts that assist us in creating tables. DW will be placed in an ORACLE database that will store all the facts and dimensions. Each dimension and segment contains an ID that associates it with the fact table; it also contains a description of the dimension. Within the same dimension may have a hierarchy where some dimensions are included in another.

Each dimension has an ID as primary key and each key ID is set as secondary key in the fact table. The star scheme is chosen for its simplicity and for the higher performance that it offers [5].

Star scheme gives us the possibility of fewer links between tables. Our scheme consists of the fact table in the center and 27 dimensions and segments.

C. Populating the DW

In our study to populate the DW are written several scripts in the SQL language. These scripts are mostly SELECT, JOIN tables of different operating systems, where the necessary information is selected and finally INSERT scripts populate the fact table.

Also the dimension tables are populated with the necessary information to different dimensions. Populating the DW will be done daily and these scripts will be scheduled to be executed every night and with sales information for a given day. To make our test the DW will be populated with sales of the 2010 and 2011, but the data of 2010 will be partly because the center is only open in March 2010.

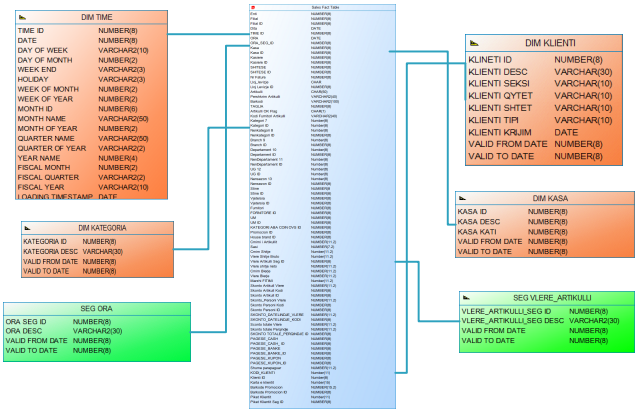


Figure 3. The scheme of DW constructed

D. Extraction of data from the DW

To extract data from the DW and create reports in a simple way, a sowfware will be used basen on our database. Microstrategy is a free soft for a limit of 25 users and a limit of 1 CPU. The program is one of the business intelligence software more appreciating and more easy to use.

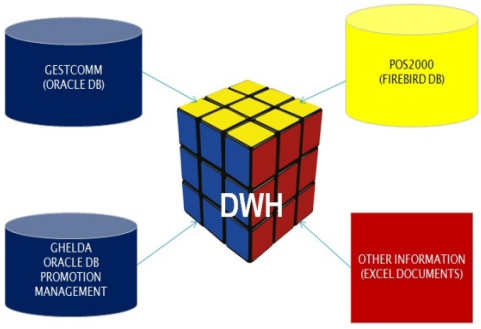


Figure 4.The sources of information from which the DW get

With the construction of this scheme, we managed to have reports on the many dimensions that before we had not and are required scripts and endless mergers tables to achieve. Now the reports are created in a very simple way and can be analyzed sales according to many parameters. Some of the data are calculated in advance before they pass in DW and this increases its efficiency. Now we can take sales reports for clients, their age, different categories of goods, the selling points used, many dimensions and segments help us in this detailed analysis.

The lackness that Microsoft has for BI is risolved by Microstrategy [6]. Microstrategy application is used to make possible to simplify the reporting according to the dimensions [7]. This application is installed on a Windows Server 2003 operating system and is connected to the server where it is located DW. The application has several modules: Desktop, Web and Mobile. There are possible configurations of levels two and three, at the third level there is an OLAP server between the user and DW.

IV. The Evaluation of Performance of Systems

To evaluate the performance in time of the reply of two systems: DW and transactional systems a query is used, that will show us the goods sold to a certain category. These jave not been in the promotion.

Query that will perform to DW will be performed to the fact table being associated with the table of dimension category and the dimension of time. The query will be performed for time periods: 3, 6, 9, 12, 15, 18 and 21 months. The database is populated with 480'000 rows belonging to sales. The time for the system response time for these periods will be measurement by *Toad application* of Oracle. This application serves as an interface using an Oracle database and provides

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the opportunity to make a query in the database in a very simple way. Figs.5 and 6 have the results for the time needed to get results from both systems.

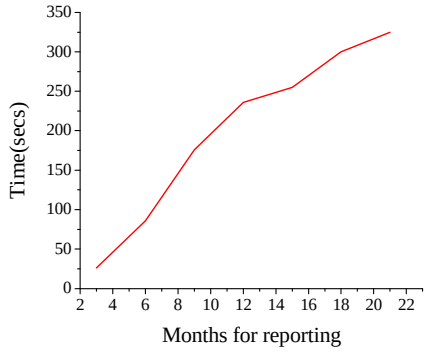


Figure 5. Time required for obatin data from transactional systems

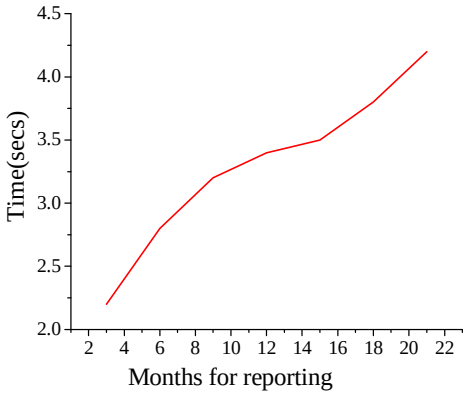


Figure 6. Time required for obatin data from a DW

As we see, results that in the case of the querie performed to transactional systems the connection between manës tables and the time that we need to get results is far greater than in the case of DW, which is structured for reporting and the fact table is populated by many sources. The collection is performed only in the fact table and time is reduced drastically. Indexation used in the relacion database is normal while in the case of DW is a bitmap that offers a better performance in the case when reading operacion to the database are much more frequent than writing operations [8], [9].

Another feature to note is that by increasing the number of rows in transactional systems is significantly increase the time and in the case of DW, by increasing rows the time grows very little. So we have a better performance of this system for reporting.

To perform a study on a database with many records should have more than 500,000 records, but in our sistem we have only 500,000 records so we will add artificially 4.5 million rows to evaluate the performance of these systemes. For add these rows will use the same INSERT scripts that we used to populate DW. The performance of ROLAP (Relational Online Analytical Processing) and HOLAP (Hybrid Online Analytical Processing technology) will be compared.ROLAP technology performs SQL query to DW while HOLAP uses intelligent cubes that are created and store in OLAP server [7]. By using Microstrategy BI will be built several cubes to dimension data.

In our simulation is used the technology. This is a technology that takes the best from both forerunner methods that were MOLAP (Multidimensional Online) and ROLAP. The first had a better performance and the second a better scalability. This technology store these cubes in memory and as a file on disk of the intelligent server.But there is a change from a simple cache. You can make different query on the cube and obtain results much more quickly than would be from DW through ROLAP technology. Scalability is good and there is no limit on the size of intelligent cubes. Whenever the server is off and on for one reason or another the file is saved to disk and reloaded in memory and the processing of a query is done in real time without DW. If the data that you want to take are not in the cube created then is used Dynamic Sourcing, that is a tool that depending on the data you wants to receive, allows to query against DW by using ROLAP or against cubes by using MOLAP. Will use 5 different data sets which contain from 500,000 to 4'500'000 rows, Tabela 1. Will create two intelligent cubes: a) a first intelligent cube consists of three dimensions : customer, category and time, b) second cube contains dimensions: cash drawers, categories and time. We have two types of architecture with three levels and four levels. For our study the architecture will be with three levels: desktop, intelligent server and data warehouse DW. Architecture is shown in Fig.7. Also architecture is shown with 4 levels where there is a web server between OLAP server and browser to that which would be possible to take existing reports and creating new.

Once the cubes are created we will perform there query which will be executed once to DW using ROLAP technology and once against cubes who use HOLAP.

Query 1 will serve to get sales for some categories grouped by sales items scattered on floors and weeks of the year.

Query 2 will serve us to make purchases of all customers for certain categories of items grouped by months.

Query 3 will serve us to obtain sales data for a client at all times, grouped by category and month of the year and its score.

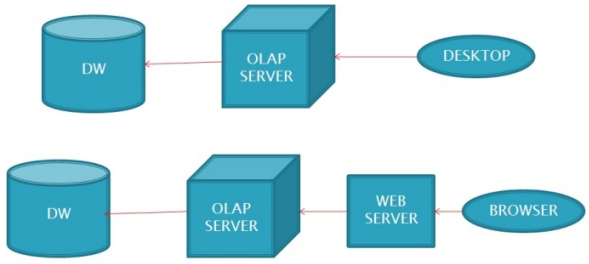


Figure 7. Architecture with three and four level

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I. THE SET OF DATA USED

Set	Number of records
1	500’000
2	1’500’000
3	2’500’000
4	3’500’000
5	4’500’000

Will measure the time response for each query and the data are summarized in the graphs. The measurement is performed ten times for each query in orden to obatin an accurate an accurate assignement of time. Will take the average value of the measured times.

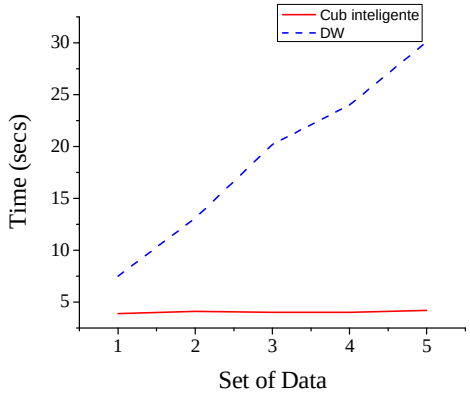


Figure 8. The response time for query 1

Intelligent cubes offer a much beter performance than ROLAP that uses query to DW. The only disadvantage is the additional memory this cubes occupy, but in this case the compression used is good and the data are stored in less memory. Cubes are stored in memory and on disk and whenever the server OLAP is start the cube is loaded from disk into memory. However if the data in the DW are updated the cube will recalculated and this takes time.But this procedure may be scheduled during night after populating the DW and during and during the day, the users can receive different reports in real time with very little delay. In our case the maximum delay goes about one minute but for a company that makes about one million rows in day, the need for intelligent cubes becomes neccessity. So, there are there different processing time for the intelligent cubes that do not depend on the query to the cube but more by the appearance of the report and rows that are to be published. In query 2 there are more lines to be published, and time of reciving the report becomes big, but this fact do not affect because the time is the same for both techniques studied. If the data that we want are not in the cube then we created a query can be done to DW by using ROLAP.

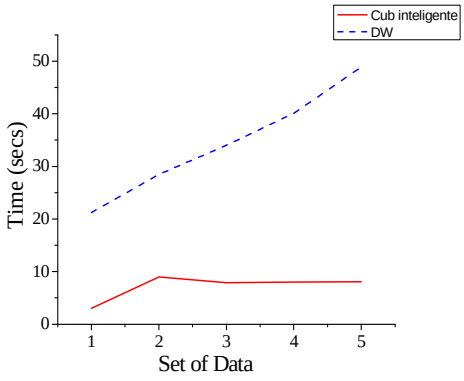


Figure 9. The response time for query 2

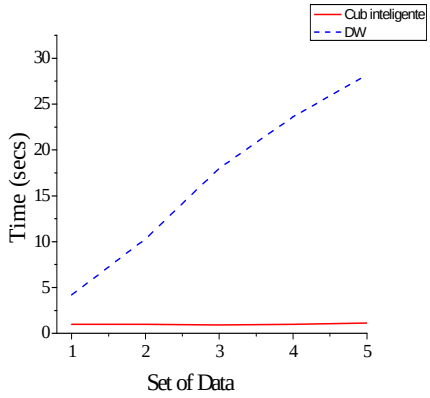


Figure 10. The response time for query 3

Automatically exists techniques called Dynamic Source which allows receiving data from the cubes if they exist, and to DW if they are not automatically set to the source data. The memory that cubes occupying is due to compression and was 1.36 Mb for the cube Client-Category-Time and 2.15 Mb for cube Cash drawer-Time Category.

The safety in accessing the program is integrated to the security of Active Directory in Windows. So, each user uses the same way with the same authentication credentials as at the AD. By Microstrategy can be determined which reports can see any user but the security filters can determine till what level can filter each user and what attributes can see.

To make very simple receiving reports will make possible the receipt by WEB reports by managers and administrators who need them. To make this possible we will use an architecture with four levels, by introducing a web server between the user and the DW, Fig.7. There is created a user for each manager and the authentication way is integrated with the authentication mode to Windows Active Directory. So they do not need to use additional credentials but can use what they use to access their profiles at the domain. When Microstrategy Web is accessible they have all the functionality they need to

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

create reports and to receive reports that already are created. Reports can be created very simply by drag and drop method. Each user can have access to one or several projects. Also we can determine the rights of each user who accesses the application.

CONCLUSIONS

The need for multidimensional data analysis as a support for decision in business has been dominant in recent years. The OLTP technology is not designed for this purpose and therefore OLAP technology was designed as a solucion.

One way to build cubes without OLAP is to write SQL query that extract the desire information from relational databases, which would be equivalent to data obtained from operations OLAP. This analysis is done in the first simulation when is compared the performance of query against the database transactional and DW. The performance is unacceptable if the database is large with lots of links between tables in organizations that hold data for several years. Unions and gatherings among many tables that in needed degrade the performance as we saw in the first test. The star scheme of DW increases the performance due to the the small number of connections that have to do to get a report. OLAP technology is optimized for these requirements to the database and therefore give a small time response. A fundamental advantage of OLAP tools is that the user gets a multidimensional information and the reporting is flexible.The data are analyzed from many dimensions and the analysis is complete. OLAP is very flexible with columns and rows, and it is possible to report in more than two dimension. The analysis for small organizations with a limited database might not need all the capacity of OLAP tools. As we saw in the first test the number of rows in the database was about 500,000 for 1 year and a half sales in the shopping center and the query to DW without using OLAP technology was only 4 seconds. But for a company that performs many transactions per day and may have at its database hundreds of thousands of rows on one day is necessary OLAP technology.

In our simulations we made a comparison of two technologies: ROLAP that performs query against DW and HOLAP which uses intelligent cubes.

Our analysis compares the efficiency of these intelligent cubes that reduce drastically the response time of the system. They also use a very good compression and the space occupied in memory is small. As the cubes are stored on the OLAP server, which means that will take reports even if the server where the database is hosted is down. Generally the usage of intelligent cubes when databases are large increases the efficiency, the performance and allows to have reports at any time even with the disadvantage of a memory occupied larger.

One of the limitations of our project was the necessity of using a single processor from 4 available because of the limitations that give us the version of Microstrategy used. Also the number of rows was limited to 4.5 million. Also the way of data entry in database DW was through SQL scripts and no ETL (Extract, Transform and Load).

In our study is not using the clustering methods that will increase the performance of the query against DW.

An improvement of the project would be through an automation tool ETL of the process of populating the DW that will give us a greater flexibility than the use of SQL scripts that we used in our study. Another area of study would be interesting the algorithms that provide the Data Mining in BI that is not take in consideration in our study.

REFERENCES

[1] M. Romm, Introduction to Data Warehousing, San Diego SQL User Group

[2] C.W.Holsapple, and A.B.Whinston, Decision support systems, New York:West publishing Company, 2000.

[3] R. Elmasri and S. Navathe. Fundamentals of Database Systems. Addison-Wesley,2004.

[4] S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. SIGMOD Rec., 26(1):65–74,.

[5] W.H.Inmon, Building the Data Warehouse, 3rd Edition, John Wiley, Chap.2,p. 36, 2002

[6] SUGI 24: The Art of Designing HOLAP Databases 2011

[7] MicroStrategy. The Case for Relational OLAP. 2010 . https://www.microstrategy.com/Solutions/5Styles/olap analysis.asp.

[8] P. E. O'Neil and D. Quass. Improved Query Performance with Variant Indexes. In Joan Peckham, editor, SIGMOD Conference, pages 38–49. ACM Press, 1997.

[9] K.Wu, E. J. Otoo, and A. Shoshani. A Performance Comparison of Bitmap Indexes. In CIKM '01: Proceedings of the tenth international conference on Information and knowledge management, pages 559–561, New York, NY, USA, 2001. ACM.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fingerprint Feature Extraction and Identification using Direction Oriented Matrix with Color Band

T. Vidhya
Dept. of Information and Communication Engineering
Sri Venkateswara College of Engineering
Sriperumbudur, India
vidhya1188@gmail.com

T.K. Thivakaran
Dept. of Information and Communication Engineering
Sri Venkateswara College of Engineering
Sriperumbudur, India
tkt@svce.ac.in

Abstract—**Fingerprint recognition is the method of biometric authentication that uses pattern recognition techniques based on the high resolution fingerprint images. Fingerprints have several advantages over other biometrics such as the following: high universality, distinctiveness, permanence and performance, easy collectable and wide acceptability. The fingerprint image is made up of pattern of ridges and valleys; they are replica of the human fingertips. The fingerprint image represents a system of oriented texture and has very rich structural information within the image. A new algorithm to extract the fingerprint features using direction matrix is proposed. The work flow is to transform the fingerprint image into 16X16 matrix by computing the orientation field using gradient information. The feature matrix is used to identify the features, in this regard, a color band information is derived which follows a different pattern for different feature based fingerprint images. The algorithm was implemented in MATLAB environment using the images from FVC databases.**

Keywords-orientation; direction matrix; pattern; identification

I. INTRODUCTION

The most popular and oldest form of biometric verification is fingerprinting [1]. Fingerprints are one of the most mature biometric identifiers and are considered legitimate identity proofs of evidence in courts of law all over the world. Therefore fingerprints are used in forensic divisions worldwide for criminal investigations. More recently, an increasing number of civilian and commercial applications are either using or actively considering fingerprint-based identification. The reason for the use of fingerprint-based identification in numerous applications is the better understanding of fingerprints as well as its superior demonstrated matching performance than any other existing biometric technology.

Fingerprint identification, known as dactyloscopy, or hand print identification, is the process of comparing two instances of friction ridge skin impressions from human fingers, the palm of the hand or even toes, to determine whether these impressions could have come from the same individual. The flexibility of friction ridge skin means that no two finger or palm prints are ever exactly alike in every detail. Fingerprint surface is made up of number of ridges and valleys interconnected. These ridges and valleys form a regular pattern with certain discriminations. Fingerprints are thus easily classified as there are four different basic shapes of pattern — arches, loops, whorls and composites — which are then subdivided according to things like the numbers of ridges between certain points in the pattern. The issue of choosing the features to be extracted should be guided by the following concerns: (1) The features should carry enough information about the image and should not require any domain-specific knowledge for their extraction, (2) The features should be easy to compute in order for the approach to be feasible for a large image collection and rapid retrieval, and (3) the features should relate well with the human perceptual characteristics since users will finally determine the suitability of the retrieved images.

The popular fingerprint representation schemes have been evolved from intuitive system design tailored for fingerprint experts who visually match fingerprints. These schemes are either predominantly local (e.g., minutiae-based fingerprint matching systems [2,3]) or exclusively global (fingerprint classification based on Henry system [4,5]). Furthermore, the conventionally used exclusively global and local approaches to fingerprint representations tend to restrict either the scope of their application or do not easily lend themselves to indexing mechanisms. The minutiae based automatic identification techniques first locate the minutiae points and then match their relative placements in a given finger and the stored template [2]. The basic idea of a direct gray-scale minutiae extraction technique [3] is to track the ridge lines in the gray-scale image, by sailing according to the local orientation of the ridge pattern. From the mathematical point of view, a ridge line is defined as the set of points that are local maxima along one direction. The ridge line extraction algorithm attempts to locate, at each step, a local maximum relative to a section orthogonal to the ridge direction. By connecting the consecutive maxima, a polygon approximation of the ridge line can be obtained. The ridge line algorithm [4] simultaneously tracks a central ridge and the two surrounding valleys. This is done by computing central maximum and two adjacent minimum values to locate the minutiae. The gray-scale neighbourhoods are extracted [6] from the original image after enhancement through Gabor filtering. Minutiae neighbourhoods are normalized with respect to minutiae angle and local ridge frequency.

On proceeding the survey, it is clearly noticed that the further matching of either minutiae detail or singular points

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

does not meet the greater accuracy, since the minutia details located are not clearly identified. This information emerges a thought how to extract the feature in a way to recognize easily as well the ridge lines must be enclosed in detail, which paved the way to transform the feature based fingerprint images to some other forms. The further analysis of transforming the feature into an understandable form from the human perspective derived an idea convert the feature into the color band pattern, which is discussed briefly in section III. The thought of the alternative way to extract the fingerprint feature which will overcome the problem is to have a detailed study of the ridge direction. While doing so, it is found that the ridge direction can be estimated using the orientation flow, which is presented in section II.

II. FEATURE EXTRACTION PROCESS

The feature extraction process involves the loaded fingerprint image is binarized and then the ridge line flow is estimated. The ridge line flow is converted to the matrix, which is transformed to easily understandable and recognizable color band pattern. The proposed work algorithm deals with the following steps:

(1) Load the image
(2) Binarization
(3) Orientation Field Estimate
(4) Feature Matrix
(5) Color Band Pattern

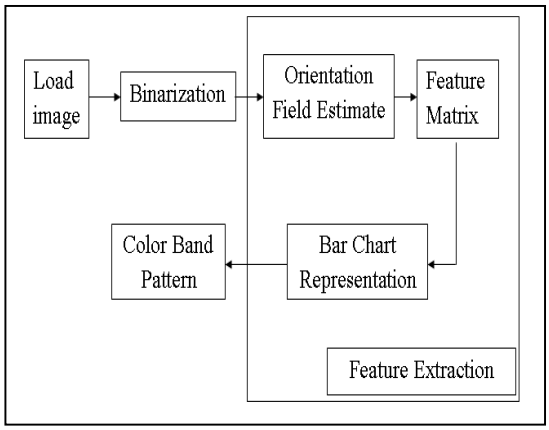
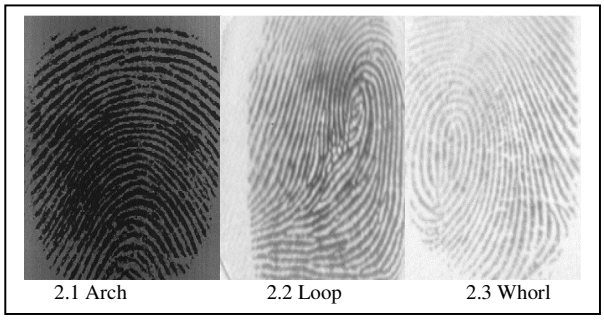


Fig. 1. Feature Extraction Process

This algorithm works in the following manner: First, the original image loaded is binarized and then the orientation field reliability is estimated. From the mathematical point of view, the orientation field is computed by calculating the horizontal and vertical gradients at each pixel using sobel operator. With the information derived from the orientation field image, the 16 X 16 feature matrix is obtained. The feature matrix is formed with the numbers say 1, 2, 3 and 4, which is obtained by shifting and rounding operation of theta values derived from the orientation image. From the feature matrix derived, the numbers and their occurrences forms different color band pattern for different feature based fingerprint images, which is determining in the identification process.

A. *Load Image*

The fingerprint images from the FVC databases: FVC 2000, FVC 2002 and FVC 2004 are loaded. Fig. 2 shows the original images from FVC databases.

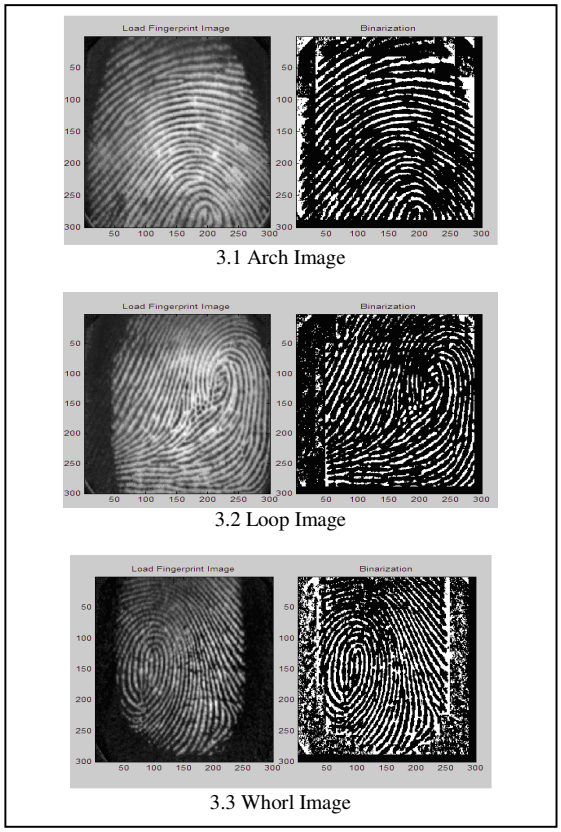


2.1 Arch 2.2 Loop 2.3 Whorl

Fig. 2. Original images from FVC Data bases

B. *Binarization*

The binarization process involves the loaded fingerprint images is divided into blocks of size sixteen. The mean value of each block is computed and thus the threshold is set, so that the ridges are shown extremely black and the background and valleys are shown white. The binarized images with that of the loaded images are depicted in Fig. 3



3.1 Arch Image

3.2 Loop Image

3.3 Whorl Image

Fig. 3 Binarization Images with Loaded Images

III. FEATURE MATRIX COMPUTATION

The binarized images depicted in Fig. 3 shows that the ridges are projected from the valleys and back ground

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

information present in the fingerprint images. The ridges are seen extreme black, still the gray information other than the ridges are also illustrated in black. Though, the binarization process helps in the analysis of the flow of ridges, it does not give clear information about the ridges as expected to extract the feature matrix. The flow of ridges must be clearly projected to identify the features present in the fingerprint images. Thus, in order to recognize the feature effectively, the orientation field estimation of the fingerprint image is computed, and transformed to easily identifiable form. The feature matrix computation involves two major processes,

(1) Orientation Field Estimation

(2) Feature Matrix Extraction

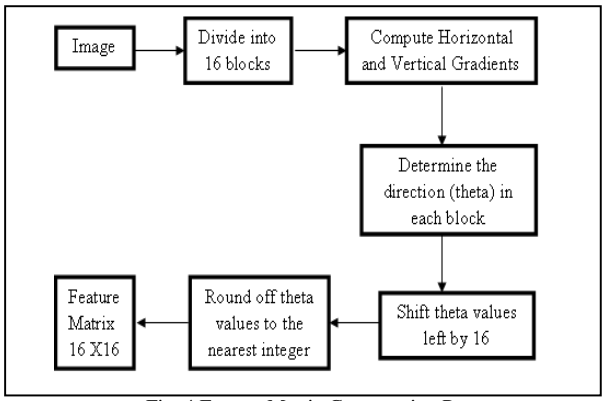


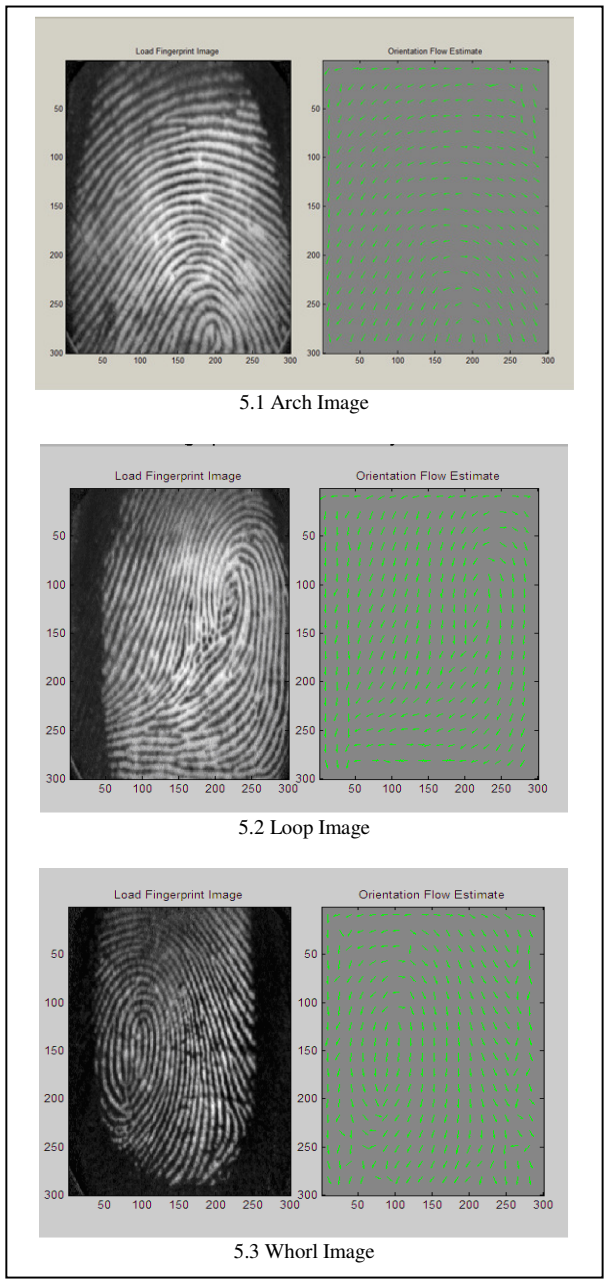
Fig. 4 Feature Matrix Computation Process

A. *Orientation Field Estimation*

The direction of ridges is represented through orientation field estimation. It exactly depicts the flow pattern of the ridges on the fingerprint surface. This flow pattern of the ridges is much enough to compute the feature matrix, which completely relies on the computation of the orientation at each block. The working rule for the estimation of orientation field is to divide the fingerprint image into number of non overlapping blocks. Subsequently, compute the horizontal and vertical gradients using simple gradient operator such as sobel. Then these gradients are averaged to compute the ridge direction information which is given by (1).

$$\theta(x,y) = \frac{1}{2}\tan^{-1}\left(\frac{2D_{xy}}{D_{xx} - D_{yy}}\right)$$
$$D_{xx} = \sum_{(x,y)\in w} D_x^2(x,y)$$
$$D_{yy} = \sum_{(x,y)\in w} D_y^2(x,y)$$
$$D_{xy} = \sum_{(x,y)\in w} D_x(x,y)\cdot D_y(x,y) \quad (1)$$

where w is the block size, and in this work it is set as sixteen. D_x and D_y are the gradient magnitude in x and y directions respectively. The Gradient information is computed using the Sobel operator. The orientation image of the corresponding loaded images are shown below,



5.1 Arch Image

5.2 Loop Image

5.3 Whorl Image

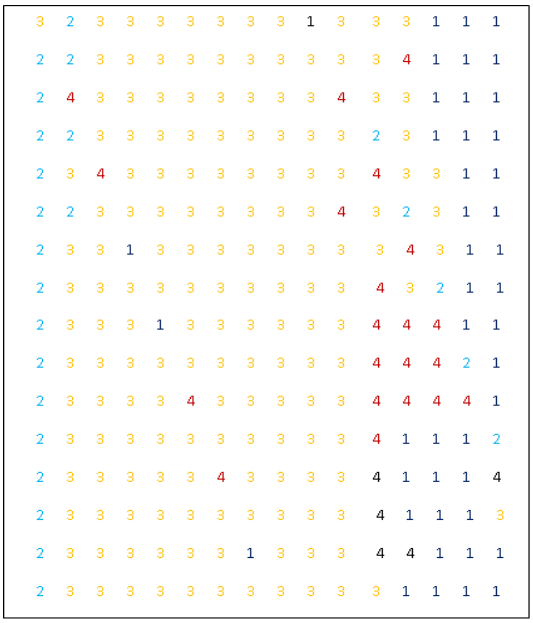
Fig. 5. Orientation Field Image for the loaded image

B. *Feature Matrix Extraction*

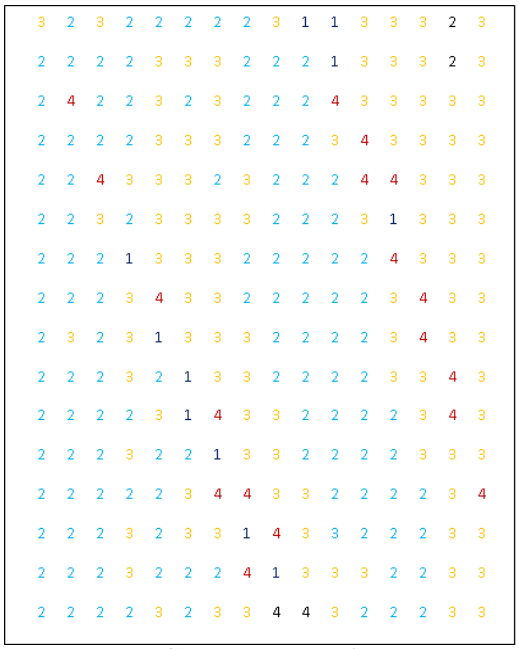
The orientation field estimation of the fingerprint images interprets the flow of ridges, which clearly identifies the different feature based fingerprint images in the human perspective. Computationally to identify the fingerprint images, feature matrix of size 16 X 16 is represented. The feature matrix is obtained from the orientation flow estimate. The theta value calculated to form the orientation images are used to design the feature matrix. The theta values signalize the direction of the flow of ridges, which possesses 256 entries in total. This is because; the image is divided into blocks of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

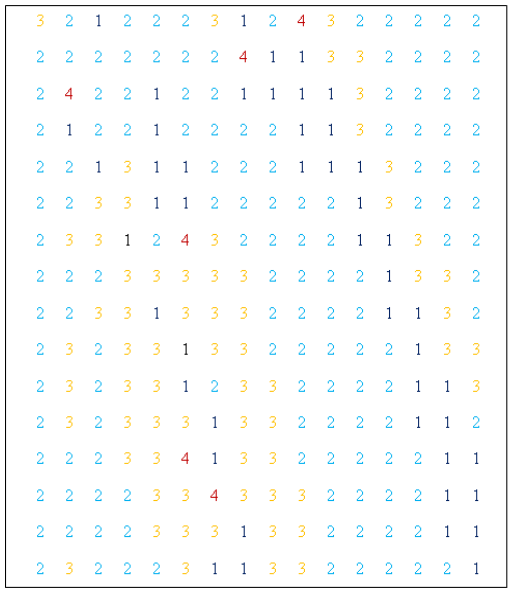
size sixteen. The theta values are rounded off to the nearest integer for the purpose of simplicity. The theta values are hence rounded off to the numbers in the range 1 to 4 values. The next step is to arrange these values into the matrix, for doing so, the values are shifted left by 16 columns. Finally, the feature matix is presented for the different featured fingerprint images. It is observed that the feature matrix is unique for the unique fingerprint sample, still forms the unique color band for distinguishable featured fingerprint images, which is briefly discussed in section IV. The feature matrix represent for the sample fingerprint images are illustrated in Fig. 6.



6.1 Arch Feature Matrix



6.2 Loop Feature Matrix



6.3 Whorl Feature Matrix

Fig. 6. Feature Matrix

From the feature matrix, it is evident that the number values 1, 2, 3 & 4, and their arrangement in the 16 X 16 matrix extract the ridge line flow. The matrix is helpful in determining the feature based fingerprint images. The feature matrix is formed with the direction information of the ridge lines in the fingerprint image, so the number of occurrences of the values in the feature matrix varies with the features. On the analysis of number of occurrences in the feature matrix, it is observed that the occurrences changes with the variation in the feature based fingerprint images. This arrives the cluster table presented in TABLE I which consists of the combination of the occurrences of the number values for the different types of fingerprint images.

TABLE I. CLUSTER TABLE

FEATURES	CLUSTER
ARCH	1,3
LOOP	1,2,3
WHORL	2,3

The number of occurrences of the values 1, 2, 3 and 4 in the feature matrix is presented in the master table for the different types of fingerprint images. The master table adds the information about the mean, standard deviation and sum of the modulus values of these number of occurrences of the values, which shows the minor differences for the variety of the fingerprint images. Thus, these values for the different flavoured fingerprint images are tabulated in the TABLE II, III and IV. Added to that, these master tables arrives the cluster and range table, which clearly identifies the different types of the fingerprint images. Thus, the feature based fingerprint images are distinguished efficiently.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II. MASTER ARCH TABLE

Image ID	1	2	3	4	Mod 1	Mod 2	Mod 3	Mod 4	Sum	Area	Ar Mean	Max
a1	89	57	95	15	78	28	66	1	173	90000	63.6154	115
a2	61	61	94	20	12	12	68	16	108	90000	63.5833	115
a3	74	55	84	43	34	36	4	41	115	90000	64	115
a4	86	42	97	31	84	4	62	8	158	90000	64	115
a5	115	88	33	20	26	80	25	16	147	90000	64	115
a6	92	74	65	25	72	34	61	6	173	90000	64	115
a7	97	86	51	22	62	84	1	14	161	90000	64	115
a8	63	82	86	25	4	10	84	6	104	90000	64	115
a9	98	61	75	22	60	12	31	14	117	90000	64	115
a10	103	92	45	16	50	72	31	0	153	90000	64	115

TABLE III. MASTER LOOP TABLE

Image ID	1	2	3	4	Mod 1	Mod 2	Mod 3	Mod 4	Sum	Area	Ar Mean	Max
l1	11	120	105	20	3	16	46	16	81	90000	64.6691	215
l2	37	115	79	25	34	26	19	6	85	90000	64.6618	215
l3	18	111	112	15	4	34	32	1	71	90000	64.6618	215
l4	120	59	66	11	16	20	58	3	97	90000	63.8977	215
l5	85	80	79	12	1	16	19	4	40	90000	63.8929	215
l6	98	121	18	19	60	14	4	9	87	90000	63.8816	215
l7	83	122	20	31	7	12	16	8	43	90000	63.8677	215
l8	46	118	79	13	26	20	19	9	74	90000	63.85	215
l9	51	126	68	11	1	4	52	3	60	93184	64.6618	215
l10	32	122	91	11	0	12	74	3	89	93184	64.6618	215

TABLE IV. MASTER WHORL TABLE

Image ID	1	2	3	4	Mod 1	Mod 2	Mod 3	Mod 4	Sum	Area	Ar Mean	Max
w1	47	125	74	10	21	6	34	6	67	90000	63.84559	143
w2	38	143	65	10	28	113	61	6	208	90000	63.84559	139
w3	41	119	65	10	10	18	61	6	95	90000	63.84559	139
w4	40	131	74	11	16	125	34	3	178	90000	63.94118	139
w5	42	130	66	18	4	126	58	4	192	90000	63.94118	139
w6	54	131	61	10	40	125	12	6	183	90000	63.93333	136
w7	48	136	66	6	16	120	58	4	198	90000	63.92593	136
w8	52	136	62	6	48	120	8	4	180	90000	63.9208	136
w9	40	136	68	12	16	120	52	4	192	90000	63.91667	122
w10	75	56	88	37	31	32	80	34	177	307200	63.8	92

IV. COLOR BAND PATTERN EXTRACTION

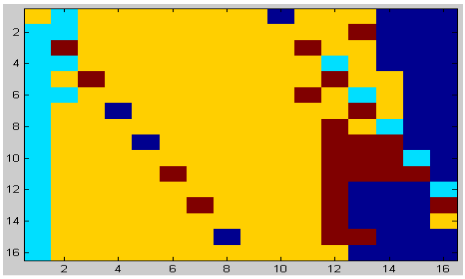
The analysis of the transformation of the feature matrix to other forms emerge an idea to the graphical charts and the color perception. Hence, the feature matrix is represented in two forms namely,

(1) Color Pattern Matrix

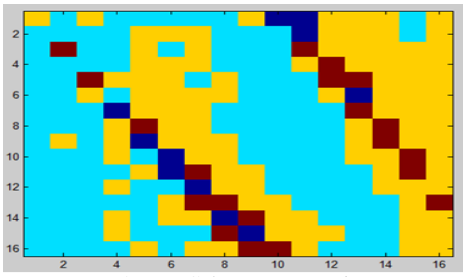
(2) Bar Chart Representation

A. *Color Pattern Matrix*

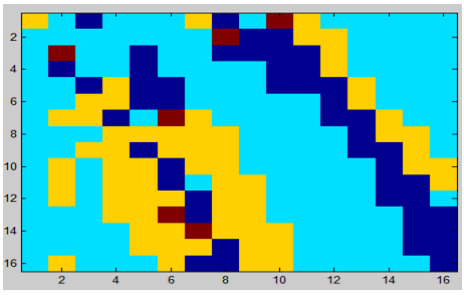
On proceeding the analysis in a way how to represent the feature matrix into easy recognizable form. The idea derived is to give the color pattern for the feature matrix obtained. The occurrence of the values in the feature matrix is replaced by the color values to obtain the color pattern matrix, which is depicted in Fig. 7.



7.1 Arch Color pattern Matrix



7.2 Loop Color Pattern Matrix

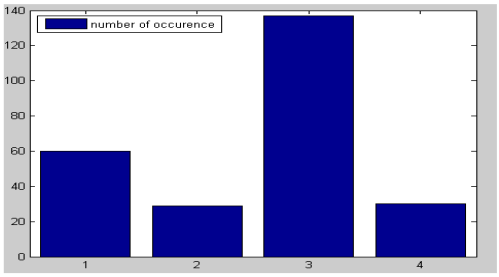


7.3 Whorl Color Pattern Matrix

Fig. 7 Color Pattern Matrix

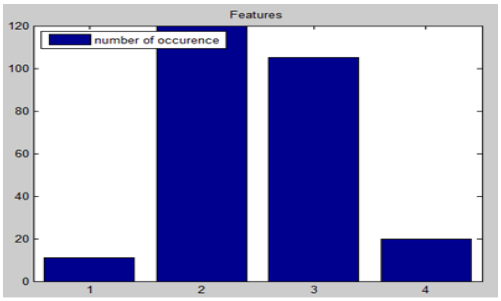
B. *Bar Chart Representaion*

The feature matrix consists of the array of numbers of range 1 to 4. From the master tables, TABLE II, III and IV, it is evident that the arrangement of these numbers differs with that of the unique fingerprint images. Thereby, the number of occurrences of these numbers varies with the distinguishable feature based fingerprint images. The numbers and their occurrences are depicted in the bar chart. The range of these occurrences is explained in the TABLE V. The bar chart proves the identity of the fingerprint images by the variation in the number of occurrences, which is shown in Fig. 8.

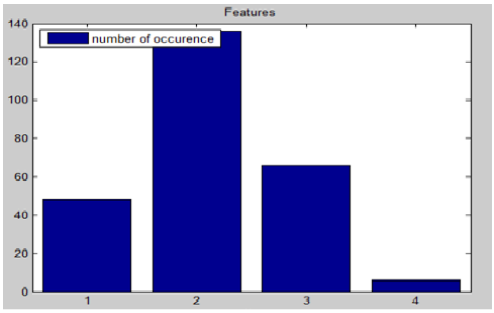


8.1 Arch Image

http://sites.google.com/site/ijcsis/
ISSN 1947-5500



8.2 Loop Image



8.3 Whorl image

Fig. 8 Bar Chart Representation

From the bar chart in Fig. 8 and the master table obtained in TABLE II, III and IV, the range values of the number of occurrences of the values are estimated in the Range Table.

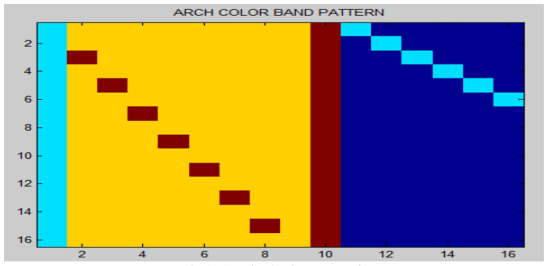
TABLE V. RANGE TABLE

FEATURES	ARCH	LOOP	WHORL
1	21-205	23-215	32-92
2	20-92	5-162	45-139
3	18-183	6-186	44-120
4	4-43	5-31	6-37

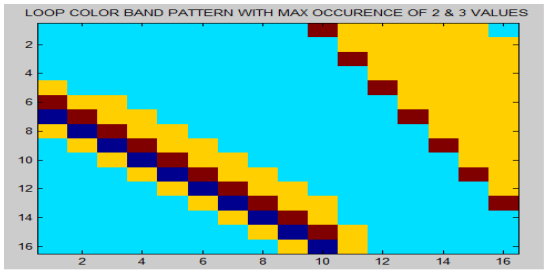
The range table gives the clear picture about the occurrences of the values in the feature matrix. The variation is seen in the range of values in the feature matrix unique for the each feature based fingerprint images.

V. RESULTS AND DISCUSSION

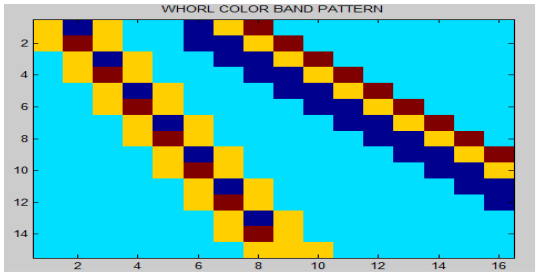
With the thorough analysis of the results obtained from the set of fingerprint images, the color band pattern depicted in Fig. 9 is derived. It is observed with the results tested with this algorithm obtained from the set of 150 fingerprint images from the standard FVC databases, the color band pattern is unique for different features of the fingerprint images. The color band is easily understandable and recognizable form to identify the different feature based fingerprint images. Added to the easily identifiable color band pattern, the bar chart obtained for the different flavoured fingerprint images and the range table and cluster table also proves the identity of the unique feature based fingerprint images.



9.1 Arch Color Band



9.2 Loop Color Band

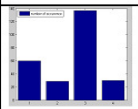
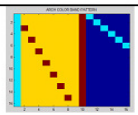
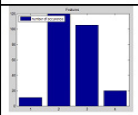
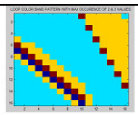
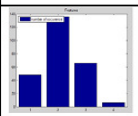
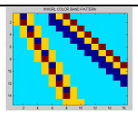
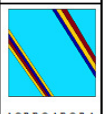


9.3 Whorl Color Band

Fig. 9 Color Band Pattern

The entire results obtained from this algorithm are tabulated in the Summarized Table, which is evident that the fingerprint feature based images are identified effectively with the color band perception which is easily recognizable and understandable. The summarized table is presented in TABLE VI.

TABLE VI. SUMMARIZED TABLE

Finger print Type	Image	Bar chart	Color Band	Pattern Band
Arch				
Loop				
Whorl				

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

VI. CONCLUSION AND FUTURE WORK

From the results and discussion, it is evident that the color band pattern identifies fingerprint features. It is concluded that the fingerprint features are extracted efficiently using direction matrix computation methodology. The proposed method is an easier way to identify the fingerprint images since it delivers the unique pattern band for the different feature based fingerprint images. Moreover, the feature matrix derivation with color band pattern method does not involve any complexity and high computational process. The future work relies on the fingerprint classification process with that of the features extracted using this algorithm.

REFERENCES

[1] A.K. Jain, S. Prabhakar," Handbook of Fingerprint Recognition", Springer, New York, 2003.

[2] A.K. Jain, L. Hong, S. Pankanti, and R. Bolle,"An Identity Authentication System using Fingerprints," *Proceedings* of *the IEEE,* Vol. 85, No. 9

[3] Maio and D. Maltoni, "Direct Gray-Scale Minutiae Detection in Fingerprints," *IEEE Trans. Pattern Anal. Machine Intell.,* Vol. 19, No. 1, pp. 27- 40, 1997.

[4] G. T. Candela, P. J. Grother, C. I. Watson,R. A. Wilkinson, and C. L. Wilson, "PCASYS:A Pattern-Level Classification Automation Systemfor Fingerprints," *NIST Tech. Report NISTIR**5647,*** August 1995.

[5] A.K. Jain, S. Prabhakar, and L. Hong, " Multichannel Approach to Fingerprint Classification" ,*IEEE Trans. Pattern Anal. and Machine Intell.,*Vol. 21, No. **4,** 1999.

[6] Maio and Maltoni (1995). Maio D. and Maltoni D., "An Efficient Approach to On-Line Fingerprint Verification," in *Proc. Int. Symp. on Artificial Intelligence (8th)*, pp. 132–138, 1995.

[7] L, Lam. Seong-Whan Lee, and Ching Y. Suen, "Thinning Methodologies-A Comprehensive Survey," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 14, No. 9, 1992.

[8] F. Leung, S.H. Leung, W.H Lau and A. Luk, "Fingerprint Recognition using Neural Network", *Proc. IEEE Workshop Neural Network for Signal Processing,* pp. 226-235, 1991.

[9] N.K. Ratha, K. Karu, S. Chen, and A.K. Jain, "A Real-Time Matching System for Large Fingerprint Databases", *IEEETrans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8,Aug. 1996, pp. 799-813

[10] A.K Jain, L. Hong, and R. Bolle, "On-Line Fingerprint Verification", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, Apr. 1997, pp. 302-313.

[11] Ross, A. Jain, and J. Reisman, "A Hybrid Fingerprint Matcher", *Proc. 16 th Int'l Conf. Pattern Recognition*, vol. 3, no.4, Aug. 2002, pp. 795-798.

[12] V. Krivec, J.A. Birchbauer, W. Marius, H. Bischof, "A Hybrid Fingerprint Matcher In Memory Constrained Environments", *Proc. 3rd Int'l Sym. Image and Signal Processing and Analysis*, vol. 2, Sept. 2003, pp. 617-620.

[13] A.K. Jain and F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters", *Pattern Recognition*, vol.24, no. 12, 1991, pp. 1167-1186.

[14] L. Hong, Y. Wan, A.K. Jain, "Fingerprint Image Enhancement: Algorithm and Performance Evaluation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, Aug. 1998, pp. 777-789.

[15] R.W. Zhou, C. Quek, G.S. Ng, "Novel Single-Pass Thinning Algorithm", *Pattern Recognition Lett.*, vol. 16, no. 12, 1995, pp. 1267-1275.

[16] M. Tico, P. Kuosmanen, "Fingerprint Matching Using an Orientation-Based Minutia Descriptor", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, Aug. 2003 ,pp. 1009-1014.

[17] S.O. Belkasim, M. Shridhar, and M. Ahmadi, "Pattern recognition with moment invariants: a comparative study and new results," Pattern Recognition, vol. 24, no. 12, pp. 1117-1138, 1991

[18] Simon Haykin, "Neural networks A comprehensive foundation", Pearson Education, Inc, 2002.

[19] A. Andre Moenssens, "Fingerprint Techniques", Inbylaw enforcement series, Sep. 1971.

[20] C. H. Teh and R T. Chin, "On image analysis by the methods of moments," IEEE Trans. On Pattern Analysis and Machine Intellligence, vol. 10, no. 4, pp. 496-523, 1988.

[21] J. Vinod VV, Ghose S. "Point Matching using Asymmetric Neural Networks". Patten Recognitin, 1993, Vol. 26, no.8, pp. 1207-1214.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Behavioural API based Virus Analysis and Detection

Sulaiman Al amro [(1,2)], Antonio Cau [(1)]
[1] *Software Technology Research Laboratory (STRL)*
De Montfort University,
Leicester, UK
{salamro; acau}@dmu.ac.uk
[2] *Information Technology (IT) Department*
Qassim University,
Buraydah, Qassim, 51452, KSA
samro@qu.edu.sa

***Abstract*—The growing number of computer viruses and the detection of zero day malware have been the concern for security researchers for a large period of time. Existing antivirus products (AVs) rely on detecting virus signatures which do not provide a full solution to the problems associated with these viruses. The use of logic formulae to model the behaviour of viruses is one of the most encouraging recent developments in virus research, which provides alternatives to classic virus detection methods. To address the limitation of traditional AVs, we proposed a virus detection system based on extracting Application Program Interface (API) calls from virus behaviours. The proposed research uses a temporal logic and behaviour-based detection mechanism to detect viruses at both user and kernel level. Interval Temporal Logic (ITL) will be used for virus specifications, properties and formulae based on the analysis of API calls representing the behaviour of computer viruses.**

Keywords-computer viruses; virus behaviour; API calls; interval temporal logic

I. INTRODUCTION

Since they first appeared, computer viruses have caused disruption to private and public organisations, governments and computer users, as they attempt to remove, modify or steal sensitive data. It is highly recommended that virus researchers should be aware of new trends, which virus writers will exploit whenever they have the opportunity. The success that attackers enjoy demonstrates that there needs to be a novel and robust detection system to prevent attacks. Therefore, a novel system is needed in order to minimise damages caused by these viruses and to defeat the new techniques used by skilful attackers.

Existing antivirus (AV) products provide detection techniques which are based on signatures that have been collected from previous seen viruses and then added to an AV database. Prior to the arriving of a virus to the system, its signature will be compared with those stored in the database and if there is a match, the virus will be detected; otherwise, the system will run normally [1]. Thus, zero day viruses will not be detected by traditional detection systems unless this new virus is received by the antivirus company and the virus signature is stored in its own database. Signature-based detection systems need databases in order to store the signatures. As the number of viruses increases every day, ever larger databases are needed to store all their signatures, so that more storage space will be needed in the near future [3,2, 14]. The large database will also affect the speed of searching for signatures, and, thus, affect the performance of the system. These disadvantages mean that the signature-based detection techniques will soon be inadequate to protect computer systems.

Behaviour-based virus detection systems have been developed recently. They do not rely on a database of signatures, but instead concentrate on the behaviour of the system. They have come to light in order to overcome the problems associated with traditional signature-based detection. The principle behind this approach is first to observe the normal behaviour of the system, after which any deviation from it will be classified as an intrusion [4]. The second is to predefine virus behaviour, so that any process which resembles virus activity can be identified as a potential virus. However, there are difficulties associated with behaviour-based detection, the greatest of which is how to define the behaviours that will detect known and novel viruses without confusing them with normal processes running in the system (known as false positives). In addition, some existing virus behaviour detection techniques rely on detecting subclasses of viruses. In general, behaviour-based detection techniques rely on identifying virus characteristics in order to detect these viruses and other viruses sharing the same characteristics in the future. One of the objectives of this research is to look into the API calls issued by computer viruses in order to specify virus behaviour that will be used in this research.

There is a growing need for behavioural specification to be used in detecting attacks, providing a robust and manageable detection technique [5, 6]. The present research proposes to build a detection technique using temporal logic specifications that have been inferred from the analysis of Windows and native API calls that represent virus behaviours. We believe that using such logical specifications and formulae will minimise the problem of the rapidly growing database of traditional AV products as well as detecting newly released viruses. A logic called Interval Temporal Logic (ITL) has been chosen to be used in this research because this logic is very suitable to describe system traces, i.e., it can be used to describe bad and good behaviours. The Tempura tool will be used to check whether a particular system trace is a good or bad behaviour.

The paper is organised as follows: Section II will provide background and related work to our framework. Section III will

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

explain the behavioural virus analysis including our API extraction mechanism. Section IV will describe how a virus can be detected at both kernel and user level using ITL formulae and Tempura. Section V will present the results.

II. BACKGROUND

It is very important to understand application program interfaces (APIs) and their features, in order to trace the behaviour of programs and to understand hidden features of malicious codes. Therefore, an outline of API calls is provided here in order to enhance understanding of this important system service.

A. Windows Application Program Interface (API) system calls

In 1995, Microsoft released Windows 95 and at the same time introduced a set of system calls known as Win32 API, which represented a 32-bit application program interface [10, 1]. The new APIs had the advantage of higher system speeds because they provided a set of optimised system operations [11]. User applications in the Windows operating system (OS) based on these API function calls are stored in dynamic link libraries (dlls) such as User32.ddl, Kernel32.dll, Advqpi.dll, and Gui32.dll, in order to gain access to system resources involving registry and network information, processes and files. Each Win32 API call has its own memory address place in the import address table (IAT) which every process in the system has and which each process will consult when it makes an API call. A Win32 API call is normally called from a process running at the user level [12], then the called API will be handled by the system and converted to its equivalent function, known as a native API call, which will be understood by the kernel of the OS. A service in the kernel will handle the requested operation and its outcome will return to the original user application that made the call [7].

The majority of systems services run in the kernel and need privileges in order to access it. Native API calls, which can be directly called by any process at the kernel level, are dealt with in the dynamic link library (ntdll.dll) in order to have the kernel provide the requester service. The complete list of kernel mode functions is stored by memory location addresses in the system service dispatch table (SSDT), which is accessed each time a native API routine is called. The parameters are then passed to the memory location and the function continues with its execution [12, 13].

As explained by [8], Windows API calls play an important role in exploiting the power of Windows, allowing virus writers to use API calls to gain more security privileges and perform malicious actions. Windows APIs issue calls to perform several actions, such as user interfaces, system services and network connections, which can be utilised for good or evil [7]. Because API calls will give a full and complete description of a particular program, the analysis of its API calls will lead directly to the understanding of its behaviour.

B. Related work

Skormin et al [8] have designed an approach that intercepts API calls while a program is running. They detect any attempt of a malware to self-replicate at run-time. Their methodology was to trace the behaviour of normal processes and analyse API calls along with their input, output arguments and the execution results. The replication of a process was modelled by the Gene of Self Replication (GSR) based upon building blocks. Each block in the GSR is considered as a portion of the self-replication process which includes seeking for files and directories, writing to files, reading from files, and closing and opening a file. This approach has detected several viruses from different classes but on the other hand, they used to hook native API calls only in the kernel. As said by [24, 13] native APIs are not fully documented that gives some viruses the ability to use some of these undocumented API to attack the system.

Alazab et al [7] have used a static analysis to track API calls using existing tools. They analyze malware to classify program executable as normal or malicious. They have used the IDA Pro [22] with their own Python program to automatically extract API calls. They had examined six groups of virus steps such as search, copy, delete, read and write. They have found that read and write files were the most API calls used by malwares to infect the program. Lists of Win32 API calls have been extracted at the user level. However, there are some viruses that might not be detected by [7] because they directly call the kernel by using native API calls as mentioned by [12, 24].

Veeramani and Rai [15] have used a statistical analysis for Windows API calls to describe the behaviour of programs. They used an automated framework for analysing and categorising executables rely on their relevant API calls. They try to increase the detection rate by using Document Class wise Frequency feature selection (DCFS) measure by getting the information related to malware from the extracted API calls. They have categorised malware into groups and the relevant APIs were extracted from these categories. DCFS based feature selection measure is used to classify the executable as malicious or benign. Their analysis and detection have been done at the user level leaving the system liable to viruses that can directly contact the kernel.

III. VIRUS BEHAVIOUR ANALYSIS

Figure 1 shows the mechanism used to analyse and extract API calls. Existing software was used to obtain information about the viruses through the following steps:

Step one: Unpack the virus.

Step two: Get the assembly code by disassembling the virus.

Step three: Extract the sequence of API calls that represent the virus behaviour.

A. Build a secure environment

In order to analyse computer viruses, a secure environment is needed to make sure that no virus can escape the system and infect other machines. In addition, some viruses will use the Internet or a local area network (LAN) to spread their malicious effects, allowing them to spread very widely indeed. Therefore, a virtual machine (VM) (Oracle VM VirtualBox) [11] was used in this research in order to secure the system.

Identify applicable sponsor/s here. *(sponsors)*

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The Linux Ubuntu operating system was used as host with Windows XP as a guest to ensure that no viruses leaked from the guest to the host, because a virus that infects one OS will not run when a different OS platform is used [1]. In some cases, viruses will use the Internet to connect to anonymous remote hosts. It is preferable not to connect to these unknown hosts, even if the virus is running on a virtual machine, so a way to prevent this is needed. However, the behaviour of viruses is the target and the Internet plays an important role in tracking these behaviours. Therefore, a fake Internet was used, allowing all the network activities in addition to allowing the tracking of virus behaviour in this research. This was achieved without causing any risk to the real Internet by installing NetKit [17], which provides a simulation of the entire Internet. NetKit was therefore installed on the host (Ubuntu) machine and then the virtual machine ran Windows XP using the fake internet.

B. *Unpacking the virus*

Packers are known as "anti-anti-virus" programs and also can be called "anti-reversing", because they exist to fight against anti-virus software as well as reverse engineering techniques. Packers are mostly used to disguise and/or compress codes. According to [7], packers are just computer programs which have the ability to restore the original executable image of a file from its encrypted and compressed one in a secondary memory location. Hence, the code might appear to do one thing, but it actually does something else, which is likely to confuse researchers.

Nowadays, computer virus writers have the benefit of using these packers to make their viruses run faster, as well as avoiding detection systems. Furthermore, the methods of packing make recognising and understanding viruses very complicated both for detection systems and analysts, because the authors can make small code modifications in order to change a signature and so avoid detection. Packing also makes analysis by researchers less easy, because to extract and understand unpacked code requires a third party tool, beside a deep and strong understanding of assembly language and the kernel, which leads to a better understanding of low level programming.

However, a number of researchers have reported the construction of tools that automatically unpack viruses such as Eureka [18], Ether [19] and Renovo [20]. The present research uses PEiD [21] to unpack the virus samples examined. PEiD is an unpacking tool that detects most common packers, cryptors and compilers for Portable Executable (PE) files. The first step was to use an interactive disassembler, IDA Pro [22], to decide whether a virus was packed or not, after which PEiD was used to indicate which packer (e.g. UPX, Upack, Xpack or PEPack) had been used. As Figure 1 shows, OllyDbg [23] was used to seek the entry point of the virus and to dump the unnecessary code. It would also save the newly unpacked virus in order to conduct a clear investigation of the malware. Our analysis shows that approximately 70% of the viruses analysed had been packed and needed to be unpacked, using the process explained above, while the remaining 30% were directly observed.

C. *Extracting the assembly code*

IDA Pro can be used to extract the assembly code from both executable (such as PE, ELF, EXE, etc.) and non-executable files and is the most practical disassembly tool [7]. It runs a static analysis and can detect whether a file is packed, as well as disassembling the code, thus providing more details and improving the understanding of the code.

IDA Pro [22] was selected as part of the API extraction mechanism used in this research because it can statically and automatically extract API calls from a file, giving an initial image of what sort of API calls the file might make. Thus, using IDA Pro allows API calls to be statically extracted and gathered, offering an important method of identifying virus behaviour. In order have more evidence about the API calls made by viruses, IDA Pro needs to be used with more than one tool that provide tracking API calls at runtime.

D. *Extracting API calls*

Viruses are just like normal programs and can be distinguished by tracking their API calls that lead to malicious actions. Therefore, this research concentrates on tracing API calls in order to understand virus behaviour. As shown in Figure 1, more than one tool [22, 25, and 26] was utilised to trace API calls in static and runtime environments. Most researchers [7, 8] rely on just one tool, which runs either statically or dynamically, but this research uses both in order to have a full understanding of what API calls have been made and when. Static analysis misses some API calls when comparing to dynamic analysis. In addition, there some Win32 and native API that appear in [25] but not in [26] and vice versa. Thus, these three tools have been used in this research to track API calls.

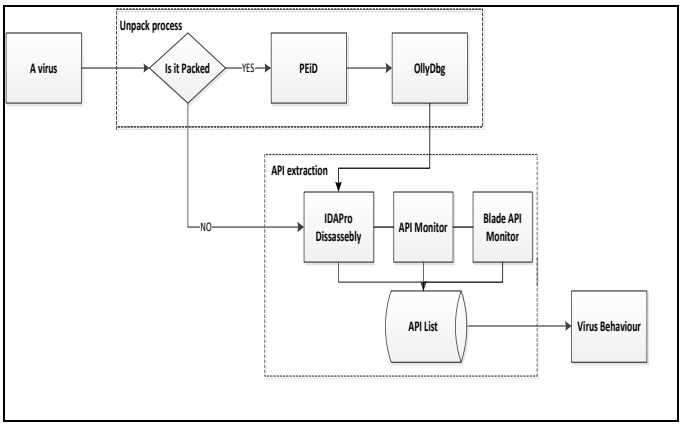


Figure 1. API extraction mechanism

The present research considers API calls to provide a way to determine whether malicious actions have been performed or not, by analysing them to understand their behaviour and to indicate whether a file contains a malicious or benign program. To do this, 283 virus samples downloaded from [29, 30] and 50 Window (XP) normal processes, such as svchost.exe and iexplore.exe, were examined to discover what sort of API calls malicious programs use in order to perform their actions.

The research began with the assumption that a virus must read from and write to a file, as [9, 8, 7] explain, in order to infect a

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

computer, to replicate itself, to infect other files and to spread throughout the world. More precisely, the following five steps are considered to represent virus steps in a behaviour:

1) Find to infect: In order to infect, a virus needs to find a file or to retrieve the contents of a directory in which to write its malicious code. This research has concentrated on three types of computer virus, listed in Table I: those which overwrite existing files, known as overwriting viruses, those which can be attached to existing files, known as parasitic viruses, and those which create a file resembling a known one, known as companion viruses.

After analysing the API calls issued by a group of computer viruses related to the three types explained above, it has been considered that 'find to infect' as the first step in the behaviour, addressing its potential API function calls that relate a search to a particular file or directory.

TABLE I. VIRUS DESCRIPTIONS

Virus type	Description	Behaviour
Overwriting	A virus (V) will replace its content with an existing file (F) by overwriting it.	Read "V.exe" Open "F.exe" Write"V.exe"into "F.exe" Close ""V.exe""
Parasitic	A virus (V) will attach itself to an existing file (F) by injecting its code into F and replace its entry points.	Open "V.exe" Read "V.exe" code Open "F.exe" Inject code into "F.exe" Replace"F.exe"entry point
Companion	A virus (V) will change the name of an existing file (F) with its original name.	Read "F.exe" Rename "F.exe" as "F.ex" Rename "V.exe" as "F.exe"

2) Get information: The second category of steps in a virus behaviour observed in this research was to discover a file's attributes, to retrieve specific information regarding a file, or to retrieve information on a directory, such as path name. A virus needs to have information about a particular file or directory to infect it and to read and write to it.

3) Read and/or copy: Read and write calls are the most important API calls issued by viruses, because they give it the ability to replicate and spread. As explained by [9], there is a very narrow difference between normal and malicious behaviour in the case of system calls. Indeed, although this research has given careful consideration to distinguishing between normal and abnormal activity, there exist some legitimate processes that may look like malicious software but would never be captured by the detector used here, because they will never act exactly the same as the malware, i.e. there is always a difference, however slight. Pervious researches such as [8] and [9] have observed that normal processes will never issue system calls that have the same order as computer viruses. This means that our concept of virus behaviour has to trace system calls from the beginning to the end, having a set of system calls which have to be made in order, because normal processes are supposed never to follow the concept of replication completely.

Therefore, read and write function calls must be made in a certain order, i.e. a virus will read a file first and then write to this or another file. In addition, other previously observed API calls of another category may or may not be called, but when it comes to read and write API calls, they must be called by the file for it to be considered a virus. Copy API calls are considered to be malicious here, because some viruses will copy to or from files when they infect a system [7]. The use of 'and/or' in the category name means that a copy API call may or may not be issued by a virus.

TABLE II. API FUNCTION CALLS FOR CATEGORIES OF STEPS

Step	Virus category	API Function Calls
First	Find to infect	FindFirstStream, FindFirstFileTransacted, FindFirstStreamTransactedW, FindFirstStreamW, FindClose, FindNextFile, FindFirstName, FindFirstFileEx, FindFirstFile, FindFirstNameW, FindNextFileName, FindNextFileNameW, FindFirstNameTransactedW, FindNextStreamW, FindNextStream.
Second	Get information	GetFileAttributesEx, GetFileAttributesTransacted, GetFileAttributes, GetFileInformationByHandle, GetFileBandwidthReservation, GetCompressedFileSizeTransacted, GetFileInformationByHandleEx, GetCompressedFileSize, GetBinaryType, GetFileSizeEx, GetFileSize, GetFileType, GetTempFileName, GetTemppath, GetFinalPathNameByHanle, GetLongPathNameTransacte,d GetFullPthNameTransacted, GetFullPthName, GetLongPathName, GetShortPathName.
Third	Read and/or copy	ReadFile, ReadFileW, OpenFile, OpenFileByld, ReopenFile, CreateHardLinkTransacted, CreateHardLink,CreateSymbolicLink Transacted, CreateSymbolicLink, CopyFileEx , CopyFile, CreateFileW, CreateFile, CopyFileTransacted , CreateFileTransacted.
Fourth	Write and/or delete	ReplaceFile, WriteFile, DeleteFileTransacted, DeleteFileW, DeleteFile, CloseHandle.
Fifth	Set information	SetFileInformationByHandle, SetFileValidData, SetFileBandwidthReservation,, SetFileShortName, SetFileAttributesTransacted, SetFileApisToOEM, SetFileAttributes, SetFileApisToANSI.

4) Write and/or delete: As mentioned in the previous subsection, a file must issue write API calls in order to be

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

classified as a virus. Therefore, every read API call should be followed by a write API call, issued at any time by the same file, to be considered a virus and not to conflict with benign processes. However, as with 'copy', the delete API call is considered malicious, because some viruses will delete some files when they infect a system, as reported by [7]. It will also be optional, as the phrase and/or appears in the category name; that is, the API delete call may or may not be issued by a virus.

5) Set information: The last category of steps in a virus behaviour observed in the research is the setting of specific information regarding a file, which leads to a change in its attributes. It has been observed that after infecting a file, a virus will need to change some of the file information in order to deal with it in the future.

Therefore, five categories of steps in representative virus behaviours, reiterated in Table II, were observed in this research and were compared with API calls to determine whether a virus was present, as explained in Figure 2.

The previous five categories were found to have used API calls that could be called by a file at the user level, known as Win32 APIs. However, there is an alternative, whereby native API calls perform this function in order to provide the service requested by the kernel. Win32 APIs are converted to native API calls by the ntdll.dll process [9], in order to be understood by the kernel. For example, when a file issues a CreateFile API call, the ntdll.dll will convert it in to its native API call, NtCreateFile, then redirect it to the kernel. However, there exist some files that can call the kernel directly, avoiding the need for user level API calls [24]. These calls were observed in this research.

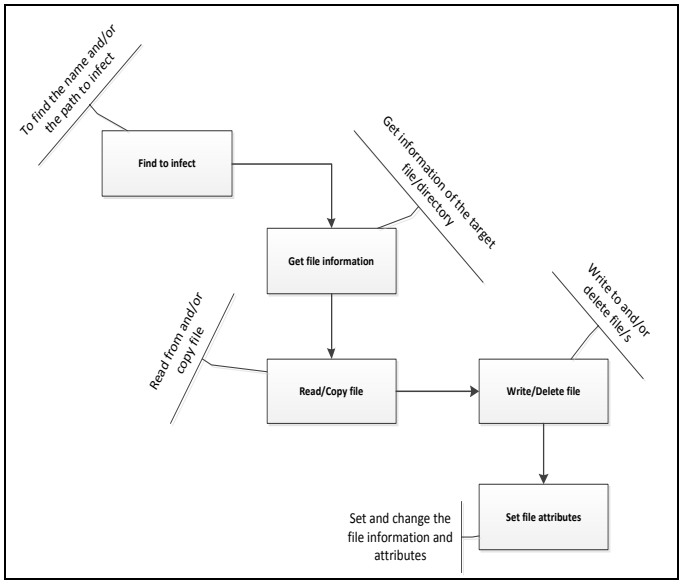


Figure 2. Five Categories

Table III shows the native API calls that can be issued by a file in order to be classified as a virus. However, these native API calls are not fully documented, as the Microsoft does not make them available [24], so API call researchers are struggling to acquire more knowledge about them. Therefore, both native and Win32 API calls need to be observed and taken into account in order for the present research to achieve good results.

TABLE III. NATIVE API FUNCTION CALLS FOR CATEGORIES OF STEPS

Step	Virus category	Native API calls
First	Find to infect	NtQueryDirectoryFile.
Second	Get information	NtQueryAttributesFile, NtQueryInformationFile.
Third	Read and/or copy	NtOpenFile, NtReadFile, NtCreateFile.
Fourth	Write and/or delete	NtWriteFile, NtDeleteFile, NtClose.
Fifth	Set information	NtSetInformationFile.

IV. VIRUS DETECTION

The observation of steps in a virus behaviour used in our system is based on the API hooking method at runtime. Hooking APIs provides the ability to intercept a set of API calls and redirect them to other functions [27, 9]. The benefit of doing so is to examine these calls in order to decide whether a virus is present or not. API hooking is done in either user or kernel mode.

A. User mode API

User mode API hooking, based on the technique of altering the IAT, redirects API calls to another place. However, Tempura will receive the API calls in order to decide whether a virus is present, as shown in Figure 3. If a virus is detected, the system will not allow it to continue making API calls and the file is terminated.

When a prototype was run in user mode only, the virus detection rate was low, because viruses are designed to evade the detection used at the user level [11]. Therefore, if no virus was detected, it was directed to the second approach and the kernel native API calls were examined.

B. Kernel mode API

The majority of computer viruses try to run at kernel level in order to gain more security levels and control of the system, which cannot be gained at the user level. At the kernel level, native API hooking does not differ from the user level, at which the SSDT can be overwritten and redefined. Therefore, any native API calls will be received at a runtime by Tempura, where they are examined for a virus.

If the user level fails to detect any suspicious API calls issued by a file, it is directed to the kernel level for further examination. If the native API calls indicate that it is suspicious (i.e. a virus), it will be terminated, while if no suspicious steps in a behaviour are detected, both API and native API calls are returned to their original file.

The disadvantage of examining API calls only in the user level is that some processes will directly contact the kernel and avoid using Win32 APIs [9, 12], allowing them to remain undetected. On the other hand, the drawback with using kernel level by itself is that unlike system calls, native APIs are not completely documented and are almost entirely hidden from view, with only handful of their functions documented in generally accessible publications [24]. This drawback makes

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the use of native APIs incomplete and liable to both false negatives and false positives, so that the system is not fully protected.

Therefore, it can be hypothesised that the use of a combined user and kernel level approach provides a better detection system and minimises the rates of false negatives and false positives. Such a system is able to examine API calls issued in the user mode and if a file is detected as a virus, no further examination is needed. If, however, it is not considered to be a virus, the detection system will examine it at the kernel level by observing its native API calls.

In order to apply this approach, a parallel execution tool is needed to run user and kernel level detection simultaneously. ITL can do this, handling both sequential and parallel composition [16] and offering user and kernel level detection at the same time. We can also make the native API calls used at the kernel level adaptable by using ITL formulae and this allows us to add more native API calls in the future.

Figure 3 shows how the system works. At the user level, API calls are extracted and then sent to AnaTempura, which examines them to see if they match the five categories of steps in a virus behaviour.

However, if the five categories are not detected in the user level API calls, Tempura examines the native API calls coming from the kernel level. This comparison is similar to the above, but concerns only those categories which have not been detected. For example, if three of the five are discovered in the first comparison, the second one considers only the undiscovered categories. Then, if kernel observation completes the set of five categories, Tempura decides that a virus has been detected and the file will not complete execution.

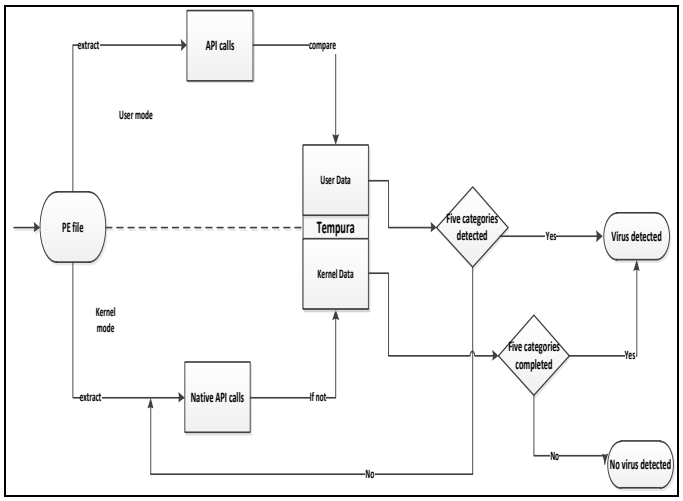


Figure 3. Virus Detection Flowchart

C. Interval Temporal Logic

ITL will be used in this research and our choice of this logic is inspired by the existence of Tempura, an executable subset of ITL that enables our virus behaviour specification to be checkable [16]. In addition, ITL is very suitable to describe system traces, i.e., it can be used to describe bad and good behaviours.

ITL is a temporal logic whose key feature is its intervals, each of which must be a non-empty, finite or infinite sequence of states σ1, σ2… A state is mapping from a set of variables *Var* to a set of values *Val*. ITL is known as a linear-time temporal logic for both infinite and finite intervals with a discrete model of time. In the finite interval it has a finite number of states and the length $|\sigma|$ of an interval is the number of these states minus one. However, a one-state interval, which is known as an empty interval, has the length zero. The sequences of states from a given system can be represented as the behaviour of this system. All behaviours of a system can be represented as the specifications of this system that can be demoted by ITL formulae using the ITL syntax and semantics.

The syntax of ITL is described in Table IV, in which μ is an integer value, a is a static variable that does not change within an interval, A is a state variable that can change within an interval, v is a static or state variable, g is a function symbol and p is a predicate symbol.

TABLE IV. THE SYNTAX OF ITL

Expressions	
$e ::=$	$\mu \mid a \mid A \mid g(e_1,\ldots,e_n) \mid \iota a{:}\, f$
Formulas	
$f ::=$	$p(e_1,\ldots,e_n) \mid \neg f \mid f_1 \wedge f_2 \mid \forall v \bullet f \mid \mathsf{skip} \mid f_1 ; f_2 \mid f^*$

The constant μ is a function without a parameter which has a fixed value, such as true, false, 1, 5. A static variable is one whose value remains unchanged in all cases within an interval (known as a global variable). On the other hand, a state variable is one that can change within an interval (known as a local variable). A function symbol can be one of several operators such as, +, -, and * (multiplication), etc. An expression of the form $\iota a{:}\, f$ is called a temporal expression. Relation symbols such as = and ≤ are used to construct atomic formulae, which will then be composed with first order connectives such as ¬, ∀ and ∃ and with skip, chop, and chopstar, which are known as temporal modalities.

D. *Informal* semantics

The beginning of an interval evaluates expressions and formulae in ITL. If there are no temporal operators in a formula, it is called a state formula. A state formula within an interval is required to hold only at the initial state of that interval. The informal semantics of the most interesting temporal constructs are defined as follows [28]:

- *skip*: is a unit interval that has a length equal to 1.
-

skip: • •

Here is a two-state Interval that has the length of 1.

- The formula $f_1; f_2$ is known to be true within an interval if it can be decomposed (chopped) into two parts, a prefix and suffix interval, such that f_1 holds for the former and f_2 for the latter, or if the interval is infinite and f_1 holds for that interval.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 4. Chop

- The formula f^* which holds for an interval is true over this interval if it can be decomposed to a finite number of intervals and the subformula is true in each of these chopped intervals.

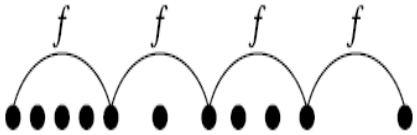


Figure 5. Chopstar

E. Derived constructs

The following constructs will be used frequently. Non-temporal derived constructs are listed in Table V and temporal derived constructs in Table VI.

TABLE V. NON-TEMPORAL DERIVED CONSTRUCTS

true	$\triangleq\ 0 = 0$	true value
false	$\triangleq\ \neg$ true	false value
$f1 \vee f2$	$\triangleq\ \neg(\neg f1 \wedge \neg f2)$	or
$f1 \supset f2$	$\triangleq\ \neg f1 \vee f2$	implies
$f1 \equiv f2$	$\triangleq\ (f1 \supset f2) \wedge (f2 \supset f1)$	is equivalent to
$\exists v\ .\ f$	$\triangleq\ \neg\forall v\ .\ \neg f$	exists

TABLE VI. TEMPORAL DERIVED CONSTRUCTS

O f	$\triangleq$ skip ; f	next
more	$\triangleq$ Otrue	non-empty interval
empty	$\triangleq\ \neg$ more	empty interval
inf	$\triangleq$ true ; false	infinite interval
isinf (f)	$\triangleq$ inf $\wedge$ f	is infinite
finite	$\triangleq\ \neg$ inf	finite interval
isfin (f)	$\triangleq$ finite $\wedge$ f	is finite
fmore	$\triangleq$ more $\wedge$ finite	non-empty finite interval
$\Diamond f$	$\triangleq$ finite ; f	sometimes
$\Box f$	$\triangleq\ \neg\Diamond\neg f$	always
Ⓦ f	$\triangleq\ \neg$ O $\neg f$	weak next
$\langle\rangle f$	$\triangleq$ f ; true	some initial subinterval
[i] f	$\triangleq\ \neg(\langle\rangle\neg f)$	all initial subinterval
◈ f	$\triangleq$ finite; f ; true	some subinterval
[a] f	$\triangleq\ \neg(◈\ \neg f)$	all subinterval

For more information on ITL syntax, semantics, derived constructs, Tempura and examples, we refer the reader to our previous paper [14], [28] and [16].

F. Virus behaviour in ITL

We have declared Cat1, Cat2, Cat3, Cat4, and Cat5 which respectively represent the lists of all API function calls for Find to infect, Get Information, Read and/or Copy, Write and/or Delete, and Set Information, as listed in Table II.

We suppose that X represents all API calls which received at runtime by Tempura. X will be received as a text representing all API calls issued by a certain PE file.

The ITL formulae for Cat1 will be as follow

$$Ucat1(X) =$$

$$inUsermode\ (X) \wedge inCat1(X). \quad (1)$$

This formula indicates that if one or more APIs calls denoted by X issued by a file in the user level, is in the list of Cat1.

The previous formula will be applicable for all the categories in the user level except Cat3 and Cat4 that represent the read and write categories respectively.

However, several rules and conditions should be considered in this research. Firstly, in order to write to an existing or new file, a virus will read and write in order, i.e. will read first and then write to the infected file. Secondly, one of the API calls (*ReadFile, ReadFileW, OpenFile, OpenFileById, and ReopenFile*) must be called in the third category and *WriteFile* API call must be called in the fourth category.

Therefore the formula of Cat3 will be as follow

$$Ucat3(X) =$$

$$inUsermode(X) \wedge inCat3(X) \wedge inRead(X). \quad (2)$$

Where Read = *(*ReadFile, ReadFileW, OpenFile, OpenFileByld, ReopenFile).

The formula for Cat4 will be

$$Ucat4(X) =$$

$$inUsermode(X) \wedge inCat4(X) \wedge X = \text{“}WriteFile\text{”}. \quad (3)$$

Because the order of read and write is very important in this research, the next formula will be applicable.

$$\Diamond Ucat3(X)\ ;\ \Diamond Ucat4(X). \quad (4)$$

It shows that the write calls must be issued sometimes ($\Diamond$) after a read call.

However, if one or more categories are not detected in the user level, then their native API calls coming from the kernel will be examined. Therefore the next formula will be used:

$$\Box(\neg Ucat1(X) \vee \neg Ucat2(X) \vee \neg Ucat3(X) \vee \neg Ucat4(X) \vee$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$\neg Ucat5(X) \equiv In\ kernel) \quad (5)$$

The previous formula indicates that if one or more categories have not been detected in the user level, they will have more examination to the kernel, in order to see if there is a call belongs to the undetected category that has been directly issued to the kernel. The same mechanism will be used to examine the native API calls coming from the kernel. Therefore, kernel level formulae will be the same as the user ones but with different predicates and variables.

V. RESULTS

30 viruses of different classes namely, Overwriting, Parasitic and Companion viruses have been observed in this experiment. We found that 29 of them can be detected by our approach. 23 of them can be detected at the user level, where 6 viruses can be detected in the kernel mode and 1 remain undetected, as shown in Table VII.

TABLE VII. DETECTED VIRUSES

Detected status	Number of viruses
User Level	(23) viruses
Kernel Level	(6) viruses
Not Detected	(1) virus

The results obtained by our research indicate that the rate of virus detection can be increased by 20%, if the two levels are examined. It also indicates that examining only the user level will leave 23% of viruses as undetected. Therefore, using a hybrid system of kernel and user level will increase the detection rate to 97%, as shown in Figure 6.

Unknown viruses will be detected by this system, if they have the similar sequence of steps of the mentioned categories. Due to the fact that this system has no database, it will not consume memory as traditional AVs. Therefore, detecting previously unseen viruses as well as minimising the memory consumption been obtained by this research.

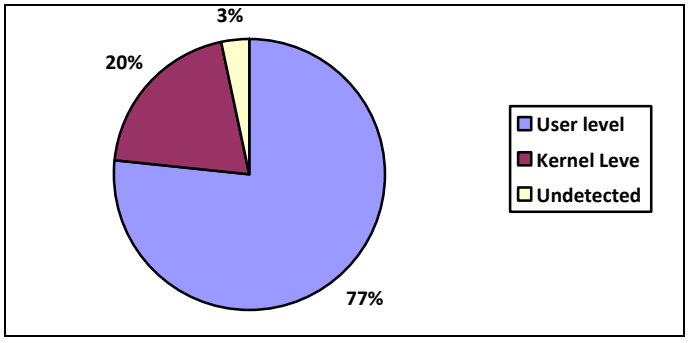


Figure 6. Virus Detection Percentage

VI. CONCLUSION AND FUTURE WORK

Most approaches that use API calls to detect computer viruses operate at either user level (Win32 APIs) or kernel level (native APIs). The problem with the former is that some applications can directly call the kernel and avoid using Win32 APIs, allowing them to remain undetected, i.e. this approach tends to give false negatives. We presented a method of detecting computer viruses in both user and kernel level by using Interval Temporal Logic. The two approaches have been used to increase the detection rate of viruses. The results indicate that the rate of virus detection can be increased by 20%, if both levels have been used. In addition to detect zero day viruses, this paper has provided a faster system by minimising the memory consumption because no database has been used in this research.

We believe that some false positives and negatives might appear when examining more and more viruses because of the exact time of switching from the user to kernel level. Therefore, our future work is to develop this research by checking both levels at the same time at statistically determined points.

REFERENCES

[1] P. Szor, The Art of Computer Virus Research and Defense, Addison-Wesley, 2005.

[2] W. Britt, S. Gopalaswamy, J. Hamilton, J. Dozier, and S. Tenaglia, "Computer Defense Using Artificial Intelligence," Proc. The 2007 spring simulation multiconference, Vol. III, ACM, pp. 378-386, June 2007.

[3] P. Harmer, P. Williams, G. Gunsch, and G. Lamont, "An Artificial Immune System Architecture for Computer Security Applications," IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, Vol. 66, IEEE Press, pp. 252-280, June 2002.

[4] M. Davis, S. Bodmer, and A. LeMasters, Hacking Exposed: Malware & Rootkits, McGraw-Hill, 2010.

[5] E. Nowicka, and M. Zawada, "Modeling Temporal Properties of Multi-event Attack Signatures in Interval Temporal Logic," Proc. IEEE / IST Workshop on Monitoring, Attack Detection and Mitigation, CiteseerX, pp. 378-386, Sep. 2006.

[6] J. Kinder, S. Katzenbeisser, C. Schallhart, and H. Veith, "Detecting Malicious Code by Model Checking," Proc. International Conference on Intrusion and Malware Detection and Vulnerability Assessment (DIMVA'05), vol. 3548, Springer, pp. 515-527, July 2005.

[7] M. Alazab, S. Venkataraman, and P. Watters, "Towards understanding malware behaviour by the extraction of API calls," IEEE 2nd Cybercrime and Trustworthy Computing Workshop (CTC-2010), pp. 52-59, July 2010.

[8] A. Skormin, D. Volynkin, I. Summerville, and J. Moronski, "Run-Time Detection of Malicious Self-Replication in Binary Executables," in Journal of Computer Security, vol. 15, no. 3, pp. 273-301, 2007.

[9] C. Seifert, R. Steenson, I. Welch, P. Komisarczuk, and B. Endicott-Popovsky, "Capture - A Behavioral Analysis Tool for Applications and Documents," 7th Annual Digital Forensic Research Workshop (DFRWS), pp. 23-30, 2007.

[10] Windows API reference. http://msdn2.microsoft.com/en-us/library/aa383749.aspx., 2011.

[11] J. Morales, "A Behavior Based Approach to Virus Detection". PhD thesis, Florida International University.

[12] G. Hoglund and J. Butler. Rootkits: subverting the Windows Kernel. Addison Wesley Professional, 2005.

[13] R. Vieler. Professional Rootkits. Wrox Press, 2007.

[14] S. Al amro, A. Cau, "Behaviour-based virus detection system using Interval Temporal Logic," Proceedings of the 6th IEEE International Conference on Risks and Security of Internet and Systems, pp.1-6, Sept. 2011.

[15] R. Veeramani, and N. Rai, "Windows API based Malware Detection and Framework Analysis," in International Journal of Scientific & Engineering Research (IJSER), vol. III, no. III, March 2012

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[16] B. Moszkowski. “Some very compositional temporal properties,” In E.-R. Olderog, editor, Programming Concepts, Methods and Calculi, volume A-56 of IFIP Transactions, IFIP, Elsevier Science B.V. (North–Holland), pp.307–326, 1994.

[17] NetKit. http://wiki.netkit.org, 2011.

[18] M. Sharif, V. Yegneswaran, H. Saidi, P. Porras, and W. Lee, “Eureka: A framework for enabling static malware analysis,” Computer Security - ESORICS, Lecture Notes in Computer Science LNCS, Springer, pp. 481-500, 2008.

[19] A. Dinaburg, P. Royal, M. Sharif, and W. Lee. “Ether: Malwareanalysis via hardware virtualization extensions,” Proceedings of the15th ACM conference on Computer and communications security, pp.51-62, 2008.

[20] M. Kang, P. Poosankam, and H. Yin, “Renovo: A hidden code extractor for packed executables,” Workshop on Rapid MalcodeWORM'07 Proceedings of the 2007 ACM workshop on Recurring malcode, pp.46-53, 2007.

[21] PEiD. http://www.peid.info/, 2011.

[22] IDA Pro Dissasember .DataRescue, “An Advanced Interactive Multi-processor Disassembler,” http://www.datarescue.com, 2011.

[23] O. Yuschuk, “OllyDbg v1.1: 32-bit assembler level analysing debugger for Microsoft Windows,” http://www.ollydbg.de. 2004.

[24] M. Russinovich, “Inside the Native API,” http://www.sysinternals.com/Information/NativeApi.html, 2005.

[25] API Monitor. http://www.rohitab.com/apimonitor, 2011.

[26] Blade API Monitor. http://www.bladeapimonitor.com/, 2011.

[27] J. Morales, P. Clarke, and Y. Deng, “Identification of file infecting viruses through detection of self-reference replication,” in Journal of Computer Virology, vol. VI., no.II, Springer. pp. 161-180, 2010.

[28] A. Cau, B. Moszkowski, and H. Zedan, “Interval Temporal Logic”, Software Technology Research Laboratory, De Montfort University, http://www.cse.dmu.ac.uk/STRL/ITL, 2007.

[29] VX Heavens (http://vx.netlux.org), 2011.

[30] Offensive Computing. http://www.offensivecomputing.net, 2011.

AUTHORS PROFILE

Mr. Sulaiman Al amro received his B.Sc degree in Computer Science from Qassim University, Buraydah (Saudi Arabia) in 2007 and M.Sc. degree in Information Technology from De Montfort University (DMU), Leicester (UK) in 2009. He is currently a PhD student at the Software Technology Research Laboratory, De Montfort University and working as a lecturer in computer science and IT department of Qassim sUniversity. His research interests are Network and System Security, Formal Methods and Computational Intelligence.

Dr. Antonio Cau gained his MSc in Computer Science from Eindhoven University of Technology (The Netherlands). He then joined Christian Albrechts University of Kiel (Germany) as a junior lecturer where he was awarded his PhD. He subsequently worked as a Research Associate on the EPSRC project `A Compositional Approach to the Specification of Systems using ITL and Tempura'. Dr. A. Cau is currently a University Senior Research Fellow at the Software Technology Research Laboratory, De Montfort University. Dr. Cau's main interests are the compositional verification and specification of critical systems using formal methods. He has developed several tools (AnaTempura, FLCheck) that can be used to accomplish this task.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

An Efficient GPU Implementation of Modified Discrete Cosine Transform Using CUDA

Massimo Panella and Luigi Basset
Dpt. of Information Engineering, Electronics and Telecommunications (DIET)
University of Rome "La Sapienza"
Via Eudossiana 18, 00184 Rome, Italy
massimo.panella@uniroma1.it

***Abstract*—A new method is presented in this paper for using general purpose programming tools of graphics processing units. It aims to calculate the modified discrete cosine transform in audio coding and compression algorithms for popular audio formats such as MP3, AAC/AC-3, and WMA. The proposed algorithm consists of matrix multiplications that are performed by the graphics processing unit. The experiments show that the proposed implementation is considerably faster than usual implementations based on early algorithms for standard hardware, so that the proposed approach can be considered for critical applications in real-time multimedia signal processing.**

Keywords-GPU computation; parallel pipelined architecture; modified discrete cosine transform; multimedia signal processing.

I. INTRODUCTION

Graphics processing units (GPUs) are very powerful tools by which the computation capability of computers is stronger than ever. Consequently, many tools have been created to program GPU for non-graphical purposes, i.e. for general purpose GPU (GP-GPU) programming [1]–[3]. The aim of this paper is to demonstrate that an algorithm implemented on GPU may work more efficiently than the same algorithm implemented on common hardware, as for example multicore CPUs or FPGA acceleration boards. However, the GPU architecture is strongly parallel and not all the applications are suitable for this implementation [4]–[6]. In spite of this, GPU overcomes CPU when there are lots of operations on big size streams.

We propose in the paper a basic framework where computational intensive applications, in particular for audio/video and multidimensional signal processing, can be efficiently pursued using GPU computation. Among all the possible applications, we focus in the following on the Modified Discrete Cosine Transform (MDCT). It is used both in subband and transform based encoding, with an analysis/synthesis system based on Time Domain Aliasing Cancellation (TDAC) [7], [8]. MDCT is designed to be performed on consecutive blocks of larger datasets and, exploiting the energy-compaction qualities of DCT, it is particularly suited to signal compression because artifacts due to block boundaries are avoided.

There are several examples in the literature concerning efficient implementations of forward and reverse MDCT for various international audio codings [9]–[12]. As a result of its advantages, the MDCT is employed in most modern lossy audio formats, including MP3, Audio Compression-3 (AC-3), Vorbis, Windows Media Audio (WMA), ATRAC, Cook, and Advanced Audio Compression (AAC). Regardless the specific application, MDCT always needs a lot of computations when the sequence to be transformed is very long. In this paper, we present a GPU implementation of a parallel and pipelined algorithm for the MDCT computation. Every iteration involves four steps and we mapped these steps into suited matrix multiplications performed by the GPU.

The proposed approach can be applied to different GPU architectures and programming environments; in this work we adopt NVIDIA™ video cards based on CUDA™ [2], [13], [14]. The GPU implementation obtained in this way is compared with a CPU implementation of MDCT, using in this case the standard programming tools. The numerical simulations will prove that GPU implementation is faster than CPU for all the considered values of frames' length.

II. PARALLEL COMPUTATION OF MDCT

Let $x(n)$ represent time domain signal samples of a real sequence. The MDCT of $x(n)$ refers to signal blocks composed of N samples (N even), which are usually windowed by a real sequence $w(n)$, $n = 0\ldots(N-1)$. The MDCT is defined as:

$$X(k) = \sum_{n=0}^{N-1} w(n)x(n+n_0)\cos\left[\frac{\pi(2n+1+N/2)(2k+1)}{2N}\right], \quad (1)$$

$$k = 0...(N/2-1),$$

where n_o is any integer determining the starting sample of the block. We outline that $X(k)$ is defined over $N/2$ samples only, because of its symmetry properties. Moreover, a definition very similar to (1) is also valid for computing the inverse transformation (IMDCT). So, what is proposed in the following for MDCT can be immediately extended also to the IMDCT.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A parallel algorithm for the computation of MDCT was proposed in [15]. The sequence $x(n)$ to be transformed is divided into N samples overlapped blocks $x^b(n)$, $b > 0$:

$$x^b(n) = x\left(n + (b-1)\frac{N}{2}\right), \quad n = 0...(N-1). \tag{2}$$

With regard to (1) it is evident that, in this case, n_0 will be an integer multiple of $N/2$. Each block is split in turn into two halves of $N/2$ samples called frames, respectively denoted as $x_1^b(n)$ and $x_2^b(n)$. Since the blocks are 50% overlapped, the second frame of block b coincides with the first frame of block b+1:

$$x_1^b(n) = x\left(n + (b-1)\frac{N}{2}\right),$$

$$x_2^b(n) = x_1^{b+1}(\mathrm{n}) = x\left(n + b\frac{N}{2}\right), \tag{3}$$

$$n = 0...(N/2-1).$$

The generic iteration of the algorithm computes the $N/2$ samples MDCT $X^b(k)$ of the current block $x^b(n)$, taking as input the $N/2$ samples of the frame $x_2^b(n)$ only. So, the algorithm is pipelined into four steps at the frame level as illustrated in Fig. 1. The four steps executed in each iteration are the following ones.

1) Computation of sequences $u_1^{b+1}(n)$ and $u_2^b(n)$, associated with the first half of the next block b+1 and the second half of the current block b, respectively:

$$u_1^{b+1}(n) = x_2^b(n)w(n)e^{-j\pi n/N}, \tag{4a}$$

$$u_2^b(n) = x_2^b(n)w(n+N/2)e^{-j\pi n/N}, \tag{4b}$$

$$n = 0...(N/2-1),$$

where $x_2^b(n)$ is used as input for both the overlapped frames.

2) Computation of Discrete Fourier Transform (DFT) of sequences $u_1^{b+1}(n)$ and $u_2^b(n)$, denoted as $U_1^{b+1}(k)$ and $U_2^b(k)$, respectively:

$$U_1^{b+1}(k) = \sum_{n=0}^{N/2-1} u_1^{b+1}(n)e^{-j2\pi nk/N}, \tag{5a}$$

$$U_2^b(\mathrm{k}) = \sum_{n=0}^{N/2-1} u_2^b(n)e^{-j2\pi nk/N}, \tag{5b}$$

$$k = 0...(N/2-1).$$

With respect to the standard definition of DFT, the exponential coefficients in (5a) and (5b) are divided by N instead of $N/2$. This means that the $N/2$ samples of $U_1^{b+1}(k)$ and $U_2^b(k)$ refer to only half of the Fourier spectrum of $u_1^{b+1}(n)$ and $u_2^b(n)$, respectively. However, no information is lost, since the sequences to be transformed are real-valued. By the way, the DFT can be computed both in CPU and GPU by using again fast computation routines, as the Fast Fourier Transform (FFT), taking into account the previous slight modification on the exponential coefficients.

3) Computation of sequences $X_1^{b+1}(k)$ and $X_2^b(k)$:

$$X_1^{b+1}(k) = U_1^{b+1}(k)e^{-j\pi(2k+1)/2N}, \tag{6a}$$

$$X_2^b(k) = U_2^b(k)e^{-j\pi(2k+1)/2N}. \tag{6b}$$

At this step, $X_1^{b+1}(k)$ is stored in memory to be used in the next iteration and $X_1^b(k)$, computed in the previous cycle, is loaded from memory for the current processing.

4) Computation of MDCT $X^b(k)$:

$$X^b(k) = \Re\left\{X_1^b(k)e^{-j\pi(2k+1)/4}\right\} + \Re\left\{X_2^b(k)e^{-j3\pi(2k+1)/4}\right\}. \tag{7}$$

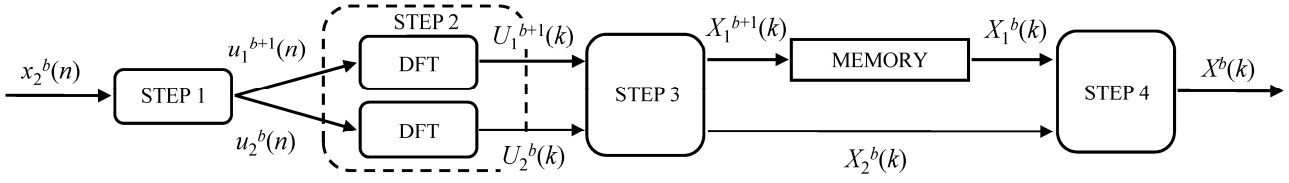


Fig. 1. Basic pipeline for MDCT computation.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. THE PROPOSED GPU IMPLEMENTATION

We implemented the previous algorithm on the GPU according to the following considerations, also summarized in Fig. 2. Firstly, the sequence to be processed is arranged in groups of frames, each consisting of $N/2$ elements; every group of M frames is stored column-wise in a $N/2$ by M matrix **F.** Without loss of generality, we will consider for the following illustrations the first M frames of $x(n)$ i.e., b=1…M:

$$\mathbf{F} = \begin{pmatrix} x_2^1(0) & \dots & x_2^M(0) \\ \dots & & \dots \\ x_2^1(N/2-1) & \dots & x_2^M(N/2-1) \end{pmatrix}. \quad (8)$$

The basic steps introduced in the previous Section are performed in parallel by the GPU array of processing elements ('stream processors' in CUDA) as illustrated in the following.

G0) The current matrix **F**, related to the frames to be processed, is loaded from system memory to GPU memory (graphic RAM).

G1) Parallel execution of the basic step 1. Every processing element of the GPU array executes a multiplication (4a) or (4b) for a frame sample; that is, either for $u_1^{b+1}(n)$ or $u_2^b(n)$ and for a given value of b and n. To this aim, we arranged the windowed coefficients of (4a) and (4b) in the columns of $N/2$ by M matrices, respectively denoted as $\mathbf{\Theta_1}$ and $\mathbf{\Theta_2}$. The generic elements of these matrices are:

$$\{\theta_1\}_{h,m} = w(h-1)e^{-j\pi(h-1)/N}, \quad (9a)$$

$$\{\theta_2\}_{h,m} = w(h-1+N/2)e^{-j\pi(h-1)/N}, \quad (9b)$$

$$h = 1\dots N/2, \quad m = 1\dots M.$$

We remark that both $\mathbf{\Theta_1}$ and $\mathbf{\Theta_2}$ are allocated once for all in the GPU memory, since they are independent of the processed frames. So, their use need not any transfer from system memory to GPU memory and vice versa.

The parallel execution of multiplications (4a) and (4b) can be obtained by instructing the GPU to perform the Hadamard product (entrywise multiplication) of **F** by $\mathbf{\Theta_1}$ and $\mathbf{\Theta_2}$:

$$\mathbf{A_1} = \mathbf{F} \circ \mathbf{\Theta_1}, \quad (10a)$$

$$\mathbf{A_2} = \mathbf{F} \circ \mathbf{\Theta_2}. \quad (10b)$$

The resulting matrices $\mathbf{A_1}$ and $\mathbf{A_2}$ are stored in the GPU memory and their columns contain the sequences $u_1^{b+1}(n)$ and $u_2^b(n)$, respectively. Evidently, there are MN multiplications to be performed by the GPU and this value is usually greater than the number of processing elements on the GPU. We can use in this case the special programming constructs provided by the development tools, in such a way the GPU interpreter/compiler can choose the optimal parallel scheduling of multiplications on its own.

G2) Parallel DFT computation of the basic step 2. The GPU processors perform in parallel the DFT of sequences $u_1^{b+1}(n)$ and $u_2^b(n)$. This is obtained by instructing the GPU to perform the FFT on the columns of matrices $\mathbf{A_1}$ and $\mathbf{A_2}$, also taking into account the previous remarks in this regard. The resulting matrices $\mathbf{B_1}$ and $\mathbf{B_2}$ are stored in the GPU memory and their columns contain the sequences $U_1^{b+1}(k)$ and $U_2^b(k)$, respectively. Even in this case, by working on the whole matrices $\mathbf{A_1}$, $\mathbf{A_2}$ and using specific instructions, the GPU will schedule the array computation of FFTs as best as it can.

G3) Parallel DFT computation of the basic step 3. Similarly to step G1, every processing element executes a multiplication (6a) or (6b) for a given DFT sample. The coefficients of (6a) and (6b) are identical and they are arranged in the columns of a $N/2$ by M matrix denoted as $\mathbf{\Omega}$. The generic element of this matrix is:

$$\{\omega\}_{h,m} = e^{-j\pi(2h-1)/2N}, \quad h = 1\dots N/2, \quad m = 1\dots M. \quad (11)$$

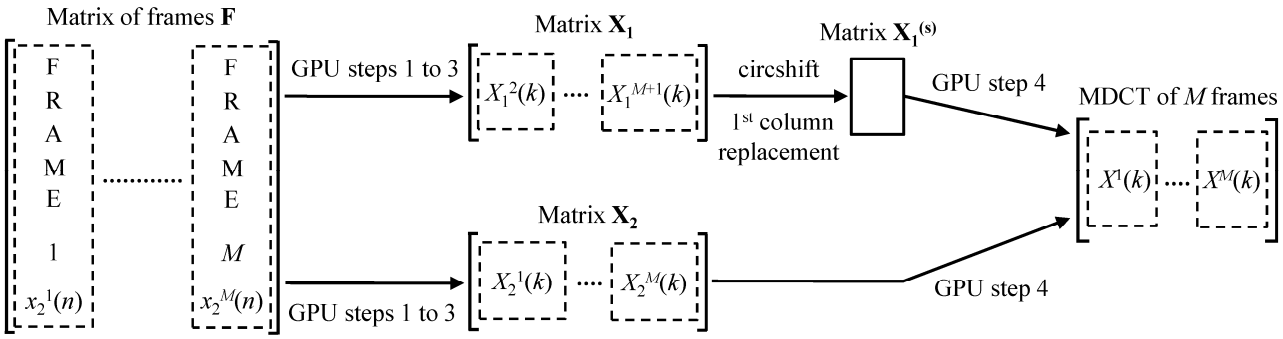


Fig. 2. GPU implementation using matrices of frames.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Also in this case, $\mathbf{\Omega}$ is allocated in the GPU memory with no transfers necessary towards system memory. The parallel execution of multiplications is once again obtained in the GPU by the entry-wise Hadamard product of the involved matrices:

$$\mathbf{X_1} = \mathbf{B_1} \circ \mathbf{\Omega}, \tag{12a}$$

$$\mathbf{X_2} = \mathbf{B_2} \circ \mathbf{\Omega}. \tag{12b}$$

The resulting matrices $\mathbf{X_1}$ and $\mathbf{X_2}$ are kept stored in the GPU memory and their columns will contain the sequences $X_1^{b+1}(k)$ and $X_2^b(k)$, respectively.

The columns of matrix $\mathbf{X_1}$ contain the sequences from $X_1^2(k)$ to $X_1^{M+1}(k)$. Conversely, the columns of matrix $\mathbf{X_2}$ contain the sequences from $X_2^1(k)$ to $X_2^M(k)$. The necessary block alignment between columns of $\mathbf{X_1}$ and $\mathbf{X_2}$ can be obtained at this stage by a right circular shift of $\mathbf{X_1}$ columns. Precisely, we determine a new matrix $\mathbf{X_1^{(s)}}$ where the columns from 2 to M will be the sequences from $X_1^2(k)$ to $X_1^M(k)$, respectively. The rightmost column of $\mathbf{X_1}$, corresponding to the sequence $X_1^{M+1}(k)$, is not discarded but stored in a GPU buffer for the successive group of frames. Consequently, the first column of $\mathbf{X_1^{(s)}}$ is obtained using the sequence stored in correspondence of the previous group of frames. In our example it corresponds evidently to the sequence $X_1^1(k)$. Thus, the resulting aligned matrix $\mathbf{X_1^{(s)}}$ will be:

$$\mathbf{X_1^{(s)}} = \begin{pmatrix} X_1^1(0) & \dots & X_1^M(0) \\ \dots & & \dots \\ X_1^1(N/2-1) & \dots & X_1^M(N/2-1) \end{pmatrix} \tag{13}$$

G4) Parallel execution of the basic step 4. Similarly to steps G1 and G3, every processing element initially runs (7) to obtain the entrywise matrix multiplications between $\mathbf{X_1^{(s)}}$, $\mathbf{X_2}$ and the related exponent matrices $\mathbf{\Phi_1}$, $\mathbf{\Phi_2}$:

$$\mathbf{Y_1} = \mathbf{X_1^{(s)}} \circ \mathbf{\Phi_1}, \tag{14a}$$

$$\mathbf{Y_2} = \mathbf{X_2} \circ \mathbf{\Phi_2}, \tag{14b}$$

where the elements of $\mathbf{\Phi_1}$ and $\mathbf{\Phi_2}$ are:

$$\{\varphi_1\}_{h,m} = e^{-j\pi(2h-1)/4}, \tag{15a}$$

$$\{\varphi_2\}_{h,m} = e^{-j3\pi(2h-1)/4}, \tag{15b}$$

$$h = 1 \dots N/2, \quad m = 1 \dots M.$$

Successively, the real parts of matrices $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are computed and added together entrywise by every processing element of the GPU:

$$\mathbf{Z} = \Re\{\mathbf{Y_1}\} + \Re\{\mathbf{Y_2}\}. \tag{16}$$

G5) The final matrix $\mathbf{Z}$, which contains in its columns the MDCT sequences $X^b(k)$, b=1…M, is stored from GPU memory to system memory.

An example of pipeline operation is illustrated in Fig. 3 by the snapshot referring to the first four time intervals, whose duration will be discussed in the next Section. In this example we consider for simplicity groups $\mathbf{F}$ consisting of only one frame, that is M = 1, so that the previous matrices consist of only one column. This may help in comparing the GPU pipeline with respect to the pipeline of the basic algorithm illustrated in Fig. 1. The GPU array is divided into four subsets of processing elements called GPU1, GPU2, GPU3, and GPU4. Each subset consists of a number of stream processors that perform in parallel the matrix operations involved in steps G1 to G4, respectively. Memory transfers in steps G0 and G5 are considered separately.

The space-time plot of Fig. 3 reveals at any time interval the data set on which each GPU subset performs its specific operations. Evidently, at the first time interval T_1 only GPU1 is fed by the frame $x_2^1(n)$, while the other GPU subsets are idle. The pipeline will be fully loaded at the fourth time interval T_4, when GPU4 will produce the MDCT $X^1(k)$ of the first block with a latency of 3 time intervals. As demonstrated successively, the computational costs of steps G1 to G4 are well balanced when using the GPU. Nevertheless, times for memory transfers (i.e., steps G0 and G5) should be considered, since they usually represent the major bottleneck for GPU computation. This is also the main reason why we arranged the sequence to be processed in group of frames, in order to reduce the critical costs due to repeated data transfers from system memory to GPU memory and vice versa.

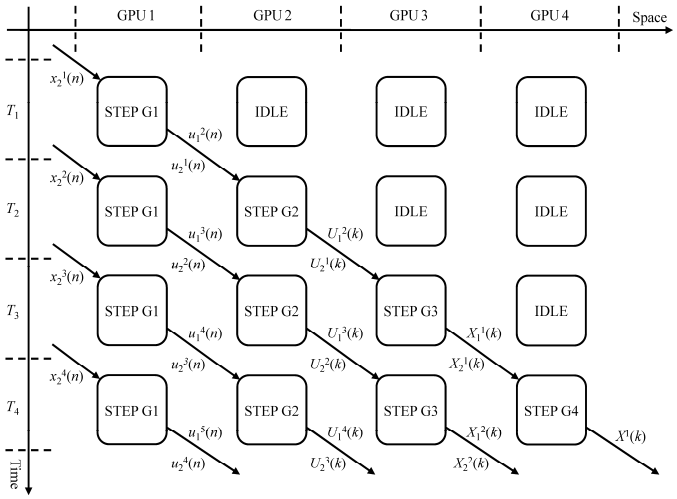


Fig. 3. A snapshot of GPU pipeline operations.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

For the sake of completeness, it is worth to mention that pipeline operations could be affine-transformed in order to improve the balance between pipeline stages. An example in this regard is reported in Fig. 4. Both throughput and latency are theoretically identical to the previous case. However, this time a GPU subset performs on the same frame all the four steps sequentially. So, this may optimize local memory transfers and distribute the load balance among all the processing elements, especially when a multicore CPU should perform some 'exogenous' tasks (mostly due to the operating system). However, we will consider in the following the more detailed implementation illustrated in Fig. 3, leaving for future investigations the optimization of pipeline scheduling.

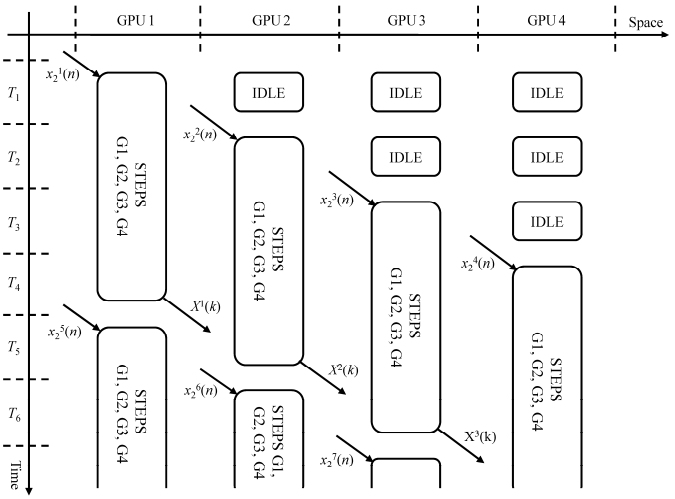


Fig. 4. An affine-transformed snapshot of GPU pipeline operations.

IV. Performance Analysis

The proposed MDCT implementation is tested on a workstation based on 4GB DDR3 RAM and a quad-core CPU AMD® Phenom II X4 945 (four K10 cores, 3 GHz clock, 45 nm technology). The system is also equipped by two NVIDIA GTX 295 dual-GPU graphics cards, for a total of four NVIDIA GT200b GPUs. Each GPU consists of 240 cores (1.242 GHz processor clock, 55 nm technology) with 896MB DDR3 RAM dedicated video memory. The software environment is based on NVIDIA CUDA C/C++ Compiler (NVCC). Both CPU and GPU systems are well balanced, with similar CMOS technology and a comparable amount of dedicated memory (about 1GB) per CPU and GPU.

By the proposed GPU implementation of MDCT, steps G1 to G4 can be run in pipeline using one GPU per step. Unlike many cases in the literature, we also consider the time spent in the pipeline for steps G0 and G5, in order to take into account data transfers between system memory and GPU memory. In a synchronous pipeline all the steps are executed in a same time interval T_{GPU}, which is the time necessary to execute the slowest step. In the same way, for the CPU implementation, steps G1 to G4 are run in parallel in the four CPU cores and data transfers G0 and G5 concern CPU and system memory only. The steps are executed in this case in the interval T_{CPU}.

As basic performance index we consider the speedup ratio between CPU processing time and GPU processing time, i.e. T_{CPU}/T_{GPU}. Furthermore, bearing in mind the algorithm proposed in the previous Section, the MDCTs of M blocks are produced together at any time interval after the step G4. So, the block rate at which the MDCT of any block is produced will be T_{GPU}/M for GPU and T_{CPU}/M for CPU. On the other hand, at any time interval the pipeline is fed by a matrix **F** consisting of $MN/2$ subsequent samples of the sequence $x(n)$ to be transformed. Hence, the *maximum* sampling rate of a sequence that can be processed by the pipeline can be $MN/(2T_{GPU})$ for GPU and $MN/(2T_{CPU})$ for CPU. We use in our tests a pseudo-random sequence consisting of $4.096 \cdot 10^6$ samples and single precision in all CPU and GPU operations. Different values of N and M are considered and the MDCT is computed 100 times for every combination of N and M, determining the mean value of T_{GPU} and T_{CPU} to be successively considered for each combination.

First of all, we noticed that the speedup is quite invariant on the problem dimension MN, which is related to the dimension $MN/2$ of matrices to be processed. This is illustrated in Fig. 5, where 8 different plots of speedup are considered. Each curve refers to a different value $N=2^i$, $i=4\ldots11$, and the speedup is computed for $M=10j$, $j=1\ldots200$. Then, all the speedups are plotted superimposed versus the problem dimension MN. It is evident that all of them have a similar behavior, being the speedup almost limited to 4 for $MN < 2.5 \cdot 10^5$ with a leap to about 12 for larger values of MN.

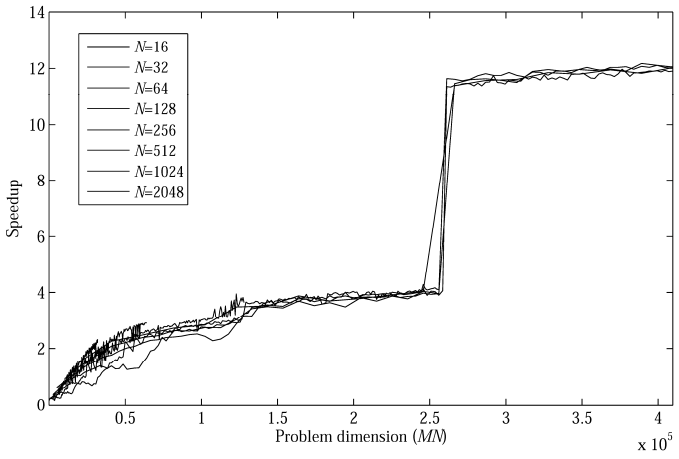


Fig. 5. Speedup factors versus problem dimension (MN).

On the other hand, a similar result can be viewed considering in Fig. 6 the speedup computed for the same values of N but plotted in this case versus M. The jump always occurs for $MN \cong 2.5 \cdot 10^5$ and the speedup can get up to 14 for $N = 2048$ and M close to 2000. We outline that the larger is N the larger is the speedup for any given value of M.

The reason why we obtain the previous behaviors depends on the time spent by GPU and CPU executing the specific operations carried out in the pipeline's steps. For instance, let us consider in Fig. 7 the time necessary to the GPU to execute each step versus the problem dimension. At any value of MN

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the slowest step determines the value of T_{GPU}. For $MN < 4 \cdot 10^4$ the slowest step is G3, where matrix multiplications are involved. Afterwards, time for memory transfers becomes dominant and the slowest steps in the GPU are G0 and G5. However, the increase of such times is quite linear on MN.

Looking at the same plots in Fig. 8, obtained in this case using the CPU, we notice that step G3 is always the slowest one in the CPU. However, there is a remarkable discontinuity by which T_{CPU} increases considerably for $MN > 2.5 \cdot 10^5$. This is the reason of the sharp steps evidenced in the speedup behaviors. Yet, the speedup is almost constant asymptotically, since T_{CPU} increases linearly for large values of MN.

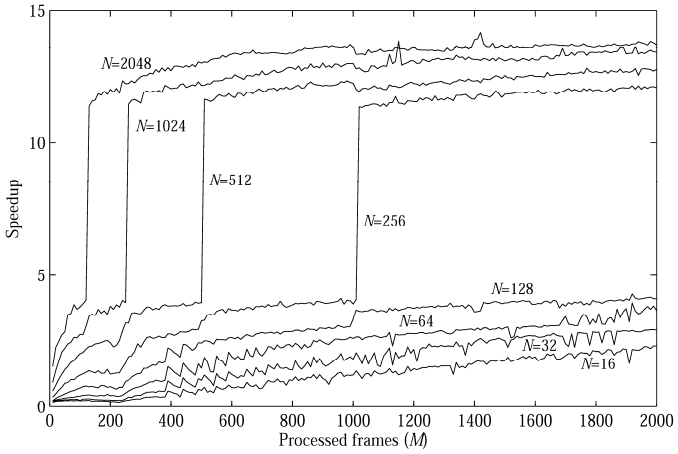


Fig. 6. Speedup factor versus M for different values of N.

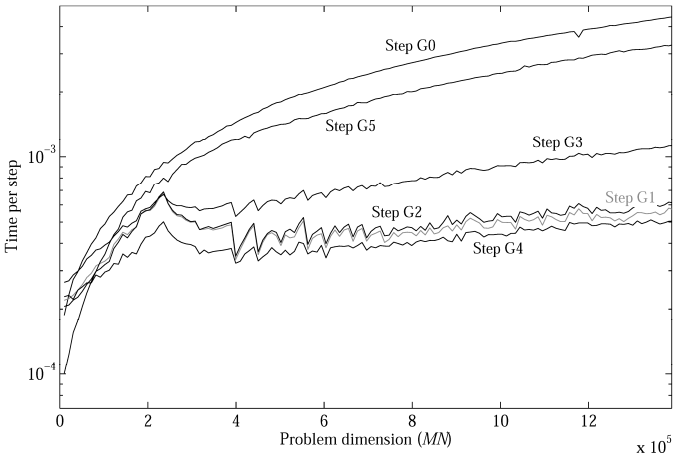


Fig. 7. GPU time per step; the gray line corresponds to step G1.

Some numerical values of speedups, block rates, and sampling rates are summarized in Table I, considering M = 1000 and several values of N. As expected in this case, both T_{CPU} and speedup step-up when $N > 256$. We remark that large values of MN are practical; for example, considering an audio sequence sampled at 44.1 KHz, a value $MN=10^6$ implies that each matrix of frames **F** will contain $MN/2=5 \cdot 10^5$ samples, that is about 11 seconds of sequence.

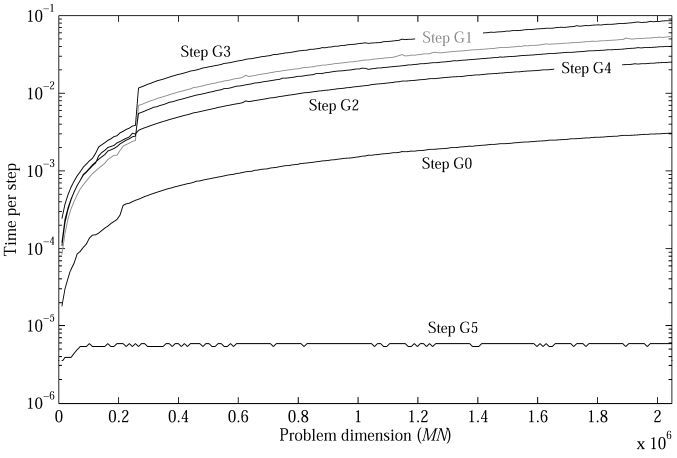


Fig. 8. CPU time per step; the gray line corresponds to step G1.

Looking at the results in Table I, the efficacy of our parallel/pipelined algorithm is confirmed, since it improves the GPU performance with respect to the CPU. In fact, a single GT200b GPU is only 8 times faster than a single K10 CPU core on common benchmarks consisting of basic functions such as LU, FFT, BLAS, 3D, etc., whereas in our application the speedup can achieve values of about 12÷14. In addition, the maximum sampling rate of sequences to be processed tends to decrease using CPU, while it can get considerable values, up to about 150 MSamples per second (or MHz), using the GPU. This reveals that the proposed algorithm is effective for the real-time processing of multichannel audio sequences even using the CPU, although in this case there can be some troubles basically related to the time sharing of CPU with other critical processes.

We can also evaluate the VLSI implementation by the well-known AT^2 product, which is quite stable by taking into account either peak or mean performances of both CPU and GPU. For instance, considering in our case the total CPU area $A_{CPU} \cong 243$ mm^2, the total GPU area $A_{GPU} \cong 1880$ mm^2, and $T_{CPU} \cong 13 T_{GPU}$, we can obtain an $A_{CPU} T^2_{CPU}$ product that is nearly 22 times larger than $A_{GPU} T^2_{GPU}$.

It is important to remark that, in this paper, we are intended to propose GPU as an interesting tool for a wider range of signal processing applications. Consequently, we are focusing on the relative efficiency of GPU against CPU for a same software framework. In other words, we rely on the efficacy of the *native* CUDA compiler for a good implementation of algorithms at machine level as, for example, for using SIMD and load-balanced multi-threading in the CPU code. In fact, we do neither consider any kind of absolute efficiency of both GPU and CPU implementations nor the relevant number of floating point operations per second (FLOPS), for one thing that many computing theorems prove that it is not generally possible to compute the number of operations in arbitrary algorithms. We should solve "the halting problem" in that case, while a mathematical proof of how the algorithm behaves would be necessary in order to compute its number of operations.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I
NUMERICAL RESULTS FOR $M = 1000$ AND DIFFERENT VALUES OF N

Block length	N	128	256	512	1024	2048
Speedup	T_{CPU}/T_{GPU}	3.319	3.904	12.252	12.982	13.604
GPU block rate	T_{GPU}/M [μs]	0.564	0.990	1.821	3.431	6.495
CPU block rate	T_{CPU}/M [μs]	1.872	3.865	22.310	44.540	88.360
GPU sampling rate	$MN/(2T_{GPU})$ [MSamples/s]	113.475	129.293	140.582	149.228	157.660
CPU sampling rate	$MN/(2T_{CPU})$ [MSamples/s]	34.188	33.118	11.475	11.495	11.589

Furthermore, in modern computers the number of FLOPS is next to irrelevant. It is more important how long those operations take, what combinations of instructions can be worked on simultaneously by the GPU or CPU, and to what extent the program can effectively make use of the sequences that are efficient to compute (e.g. combined multiply-and-add producing one answer per clock cycle after a small delay). It is also important, for efficiency, the ability of the algorithm to fit into cache.

V. CONCLUSION

In this paper we propose a parallel and pipelined algorithm for the computation of MDCT. Our method consists of various matrix multiplications that can be performed by standard hardware such as CPUs, FPGA acceleration boards and, in particular, by GPUs using the NVIDIA CUDA architecture. The numerical results show that the proposed implementation is considerably faster than usual methods. In fact, considering the obtained speedups, it is clear that our algorithm implemented on GPU may work more efficiently than the same implementation on a CPU.

The proposed algorithm is able to perform efficiently on audio sequences even using a CPU. However, what is proposed in this paper should be considered as a basic framework where further applications based on GPU computation can be developed. In fact, we are currently focusing on such multimedia applications for which the use of GPU is desirable because of big data streams and high data rates as, for instance, in the case of high quality and real-time MPEG-4 video encoding.

ACKNOWLEDGMENT

This work has been in part supported by the Italian Ministry of Education, Universities and Research (MIUR). The authors would also like to thank Dr. Daniele Mardero for his helpful contribution in developing some basic software tools that have been used in the experimental tests illustrated in the paper.

REFERENCES

[1] J.D. Owens, M. Houston, D. Luebke, S. Green, J.E. Stone and J.C. Phillips, "GPU Computing", *Proc. of the IEEE 2008*; Vol. 96, n. 5, pp. 879-899, 2008.

[2] M. Harris, "Optimizing Parallel Reduction in CUDA", *NVIDIA Developer Technology*, 2008.

[3] Wu Enhua and Liu Youquan, "Emerging technology about GPGPU", in *Proc. of IEEE Asia Pacific Conf. on Circuit and System*, pp. 618-622, Nov. 2008.

[4] R. Bordawekar, U. Bondhugula and R. Rao, "Believe it or not!: mult-core CPUs can match GPU performancefor a FLOPintensive application!", in *Proc. of Int. Conf. on Parallel Architectures and Compilation Techniques* (PACT '10), DOI: 10.1145/1854273.1854340, Sep. 2010.

[5] V.W. Lee, C. Kim, J. Chhugani et al. "Debunking the 100X GPU vs. CPU myth: an evaluation of throughput computing on CPU and GPU", in *Proc. of Annual Int. Symp. on Computer Architecture* (ISCA '10), DOI: 10.1145/1815961.1816021, Jun. 2010.

[6] R. Vuduc, A. Chandramowlishwaran, J. Choi, M. Guney and A. Shringarpure, "On the Limits of GPU Acceleration", in *Proc. of USENIX Conf. on Hot Topics in Parallelism* (HotPar'10), Jun. 2010.

[7] J.P. Princen, A.W. Johnson and A.W. Bradley, "Subband/Transform coding using filter bank designs based on time domain aliasing cancellation", in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing* (ICASSP '87), Vol. 12, pp. 2161-2164, Apr.1987.

[8] J.P. Princen and A.W. Bradley, "Analysis/Synthesis filter bank design based on time domain aliasing cancellation", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol 34, n. 5, pp.1153-1161, 1986.

[9] V. Britanak and K.R. Rao, "An efficient implementation of the forward and inverse MDCT in MPEG Audio Coding", *IEEE Signal Processing Letters*, Vol. 8, n. 2, pp. 48-51, 2001.

[10] V. Britanak, "A unified fast computation of the evenly and oddly stacked MDCT/MDST", in *Proc. of 4th EURASIP Conf. focused on Video/Image Proc. and Multimedia Comm.*, Vol. 1, pp. 233-238, Jul. 2003.

[11] V. Nikolajevic and G. Fettweis, "Computation of forward and inverse MDCT using Clenshaw's recurrence formula", *IEEE Trans. on Signal Proc.*, Vol. 51, n. 5, pp. 1439-1444, 2003.

[12] Che-Hong Chen, Bin-Da Liu and Jar-Ferr Yang, "Recursive architectures for realizing modified discrete cosine transform and its inverse", *IEEE Trans. on Circuits Syst.-II: Analog and Digital Signal Proc.*, Vol. 50, n. 1, pp. 38-45, 2003.

[13] "NVIDIA CUDA™ Programming Guide - version 4.2," NVIDIA®, available at http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf, Apr. 2012.

[14] S. Ryoo, C. Rodrigues, S. Baghsorkhi, S. Stone, D. Kirk and W. Hwu. "Optimization principles and application performance evaluation of a multithreaded GPU using CUDA", in *Proc. 13th ACM SIGPLAN*, DOI: 10.1145/1345206.1345220, Feb. 2008.

[15] N.R. Murthy and M.N.S. Swamy, "A parallel/pipelined algorithm for the computation of MDCT and IMDCT", in *Proc. of IEEE Int. Symp. on Circuits and Systems* (ISCAS '03), Vol. 4, pp. 540-543, May 2003.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

AUTHORS PROFILE

Massimo Panella received the Dr. Eng. degree in Electronic Engineering (five-year graduation with Honors) in 1998 and the Ph.D. degree in Information and Communication Engineering in 2002 from the University of Rome "La Sapienza" (Italy). Since 2001 he is an Assistant Professor (Researcher) of Circuit Theory and Computational Intelligence at the University "La Sapienza". His research interests regard circuit theory, computational intelligence and sensor networks applied to pervasive systems, multimedia, intelligent transportation systems and intelligent signal processing.

Luigi Basset received the B.Sc. degree in Electronic Engineering in 2007 and the M.Sc. degree in Telecommunication Engineering in 2010 from the University of Rome "La Sapienza" (Italy). After an internship program with Poste Italiane Spa, he currently follows the WiMAX network planning in the south-east of Italy for Linkem Spa. He is currently focusing on new base station development and he is responsible of the design of the radio access network and the transport network through microwave links.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Using Hybrid Decision Tree -Houph Transform Approach For Automatic Bank Check Processing

Heba A. Elnemr
Computer science Departement, Akhbar Elyoum Academy
Computer and systems department, Electronics Research Institute
Giza, Egypt
heba_elnemr@yahoo.com

Abstract*—one of the first steps in the realization of an automatic system of bank check processing is the automatic classification of checks and extraction of handwritten area. This paper presents a new hybrid method which couple together the statistical color histogram features, the entropy, the energy and the Houph transform to achieve the automatic classification of checks as well as the segmentation and recognition of the various information on the check. The proposed method relies on two stages. First, a two-step classification algorithm is implemented. In the first step, a decision classification tree is built using the entropy, the energy, the logo location and histogram features of colored bank checks. These features are used to classify checks into several groups. Each group may contain one or more type of checks. Therefore, in the second step the bank logo or bank name are matched against its stored template to identify the correct prototype. Second, Hough transform is utilized to detect lines in the classified checks. These lines are used as indicator to the bank check fields. A group of experiments is performed showing that the proposed technique is promising as regards classifying the bank checks and extracting the important fields in that check.

Keywords-component; automatic check processing; color histogram; houph transform; statistical features; decision tree classifier

I. INTRODUCTION

The widespread use of bank checks in daily life makes the development of check processing systems of fundamental relevance to banks and other financial institutions. Bank transactions involving checks are still increasing throughout the world in spite of the overall rapid emergence of electronic payments by credit cards [1]and [2]. there are huge volumes of handwritten bank checks that are processed manually every day. In such a manual verification, user written information including date, signature, legal and courtesy amounts present on each check has to be visually verified. Hence, much time, effort and money can be saved if this entire process of recognition, verification and data entry is done automatically using images of checks [3].

The automatic processing of a bank check involves check preprocessing, classification, extraction and recognition of handwritten or user entered information from different data fields on the check such as courtesy amount, legal amount, date, payee and signature [3]- [8]. This is a formidable task and requires efficient image processing and pattern recognition techniques.

The most common goal of automatic bank check treatment systems in previous literatures is the recognition of handwritten information. However, in order to do this, it is necessary to use a reliable and efficient process that is capable of classifying the bank checks, identifying the important areas and extracting the information, which can then be submitted to a further recognition phase. Therefore there are many reports on the extraction of filled in items of bank checks [11]-[17]. However, these methods cannot be applied directly to Arabic bank checks due to their specific properties.

In the case of Arabic bank checks, checks differ not only in background, but also in type and position of the preprinted information fields and of the information fields that must be filled in by the customer. Many different color pictures are used as background images, besides various types of stamps appear in the background. Furthermore, many other components are typically preprinted on a bank check, for instances boxes and guidelines. Finally, it must be noted that the size and the structure of a bank check can change consistently not only depending on the country, but also on the bank. Therefore, a successful system has to deal with these various forms of bank checks, colored backgrounds, drawings and printed

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

text. Thus, there is an urge to fully automate the bank check classification process before the segmentation and recognition tasks. However, there are only a few researches in automatic classification of bank checks in literatures [5], [6] and [7].

In [5] a heuristic algorithm for Arabic bank check classification, based on extracting a set of features (size, color and logo or bank name location), is proposed. The extracted features are used to build a decision table, which is used to classify the checks automatically. To increase the average recognition rate, the authors utilized genetic algorithm for the check classification [6] and [7]. The proposed technique is based on hierarchical approach. The first-level attempts to assign the input image to a restricted set of classes within the data base, while the second-level returns the best matches to each of the selected classes. But their work had several shortcomings. The proposed approach in [5] is sensitive for stamps, which reduced the recognition rate. Regarding the algorithm presented in [6],

- The authors did not take into consideration the time required in the training phase
- The authors did not take into consideration that the selected regions may be distorted by the existence of some stamps which may cause wrong classification.
- Each run with the GA produces a different set of patron regions that can be used in check type identification. Thus, human intervention is needed to select the best run.

Therefore, a technique for designing a fast and robust classifier for checks that may undergo variations of handwritten information and stamps due to individuals must be developed.

Furthermore, after check classification, it is necessary to perform the extraction operation of different handwritten fields prior to their recognition. [9] Presents an algorithm for extraction of segmented fields accomplished by means of a connectivity-based approach. For bank checks written in Bengali language [10], a template describing the locations of user-entered data is used to extract the areas of interest. A stored background pattern is then subtracted from these sub-images to eliminate the back-ground. An approach for identifying the courtesy amount string is developed in [11]. This approach involves three stages: organizing the information in blocks using connected components; identifying the string candidates; and finally deciding which string represents the amount of the check. In [12], for Arabic bank checks, two methods based on mathematical morphology and Hough transform are used for the extraction of handwritten Arabic zones of complex documents. It is found that the Hough transform based method performs better than Mathematical Morphology. A hybrid method of extraction of handwritten areas of Tunisian bank check is presented in [13]. First the bank model is recognized using the bank code, after that the color of the handwritten is detected. Finally, the extraction is carried out using mathematical morphology tool.

[14] and [15] proposed a method to extract the handwritten date, courtesy and legal amounts of Canadian bank checks. The method is based on the determination of baselines of checks, a priori information about the positions of data on checks, and a layout-driven item extraction technique.

In this paper we propose a new algorithm to classify Arabic bank checks and to extract separately filled-in items from the classified checks. The classification approach is based on a prior knowledge about color characteristics and layout of the check models.

The proposed classification system is composed of several modules. In the first module, the preprocessing algorithm explained in [5] is achieved. After that, the statistical color histogram features, the entropy, the energy are extracted. Next, morphological operations are used to detect the logo position in the processed checks. Finally, the classification technique is achieved through two stages. In the first stage, a decision tree classifier is built. The decision tree is used to classify checks into several groups. Each group may contain one or more type of checks, therefore bank logo or name is matched against its stored template. On the other hand, the extraction of the important fields is based on houph transform. The baselines are detected, and then the zones around these lines are extracted. Finally, the connected components within these zones are determined.

This paper is organized as follows. The statistical color histogram features, the entropy and the energy extraction are first presented in section 2. Section 3, describes the logo localization and extraction method, while the decision making stage is presented in section 4. The extraction of handwritten information is described in section 5. Section 6, shows the experimental results. The conclusion is finally discussed in section 7.

II. FEATURE EXTRACTION

This stage aims to classify the input check into predetermined clusters. These clusters are obtained based on color histogram of different reference check models. Arabic bank checks may differ obviously in color. Thus, these differences can be used as features to discriminate between different types of checks; this may be done by extracting statistical features from the color histogram. The statistical features extracted are

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the mean, standard deviation, and entropy which signify the important texture features of check image.

A. Statistical feature extraction

In general, any image processing and analysis applications would require a particular feature for classification /segmentation. Mainly statistical features are of more significant in pattern recognition area. A frequently used approach for texture analysis is based on statistical properties of intensity histogram. Moreover, Color is a significant feature in discriminating between different check types. This is very important as it is invariant with respect to scaling, translation and rotation of an image. The histogram plot is used to display the brightness of the colored image, showing the occurrence of pixel counts for all 256 intensity levels. The occurrence probability function of the gray level k can be simply estimated from the histogram [16], which is formulated as follows:

$$p(k) = \frac{h(k)}{\sum h(k)} \quad (1)$$

where p(k) is the probability distribution function and h(k) is the histogram function.

Entropy is the measure of the image information content, which is interpreted as the average uncertainty of the information source. Discrete entropy is the summation of the products of the probability of outcome multiplied by the log of the inverse of probability of outcome, taking into considerations all possible outcomes {1, 2, …, n} in the event {x1, x2, …, xn}, where n is the intensity level; p(i) is the probability at the gray level of i, which contains all the histogram counts. It is formulated as follows

$$E(x) = \sum_{i=0}^{G-1} p(i)\log_2(1/p(i)) = -\sum_{i=0}^{G-1} p(i)\log_2 p(i) \quad (2)$$

$$\sum_{i=1}^{G-1} p(i) = 1 \quad (3)$$

Where G is the number of gray levels.

The mean and standard deviation of pixels of the bank check image can be computed based on color histogram as follows.

$$\text{mean} = \sum_{i=0}^{G-1} i * h(i) \,/\, \sum_{i=0}^{G-1} h(i) \quad (4)$$

$$\text{Variance} = \sqrt{\frac{\sum_{i=0}^{G-1} h(i)*(i-\text{mean})^2}{\sum_{i=0}^{G-1} h(i)}} \quad (5)$$

Based on the values of these features, an attempt to classify the input check image into six basic categories is performed. Each category contains one or more of check types. Thus, logo is then used to cluster checks within each category into several groups. This will be our point of interest in the next section.

B. Logo localization

In this section we describe an approach to locate and extract bank check logos automatically; our approach is a modification of that proposed in [5]. This algorithm employs mathematical morphology. The logo location and extraction process is carried as follows. First, the colored image is thresholded. Second, three erosion iterations are applied to the thresholded check image. It was found that three erosion iterations were adequate to eliminate all the undesirable information and leave parts of the logo [5]. Hence, the largest continued black area of the eroded image is extracted and its starting and ending positions are recorded. In practice, however, because of the presence of various stamps in different locations and because of the fact that the size of the logo component is irregular as well as that some Arabic banks do not have logos and have only bank name, this is not always happens. Researchers in [5] suggested dividing the check into 3 × 3 regions, and the starting and ending positions are localized with respect to these regions. This was not adequate to remove the effect of stamps on bad localization. Therefore, in this work we suggest first to threshold input bank check at 100, to eliminate as much as possible of the stamps and handwritten effects. Also, we suggest dividing the check into 4 x 3 regions. This will reduce the influence of overlapping stamps or handwritten strokes. The extracted black area of the erosion image is thresholded. Moreover, it is obvious that logos cannot be located in the regions 1, 2, 4, 5, 7 and 8 since they are registered for text information. The logo location and extraction process is illustrated in Figure 1. Consequently, checks are divided into two subgroups with logo and without logo groups. Besides, checks that have logos are divided into several groups according to the location of the logo.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

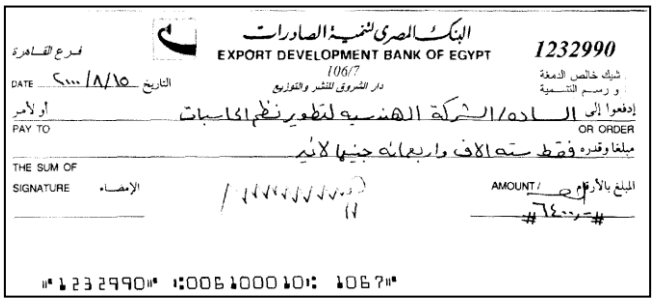


Figure a

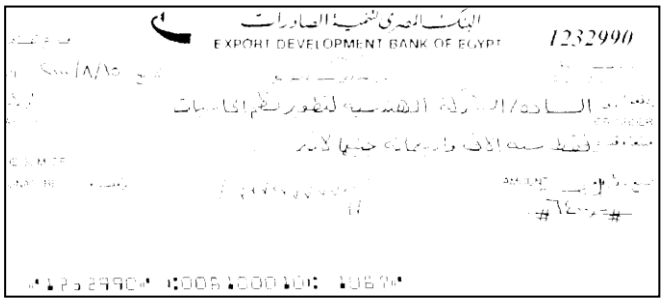


Figure b

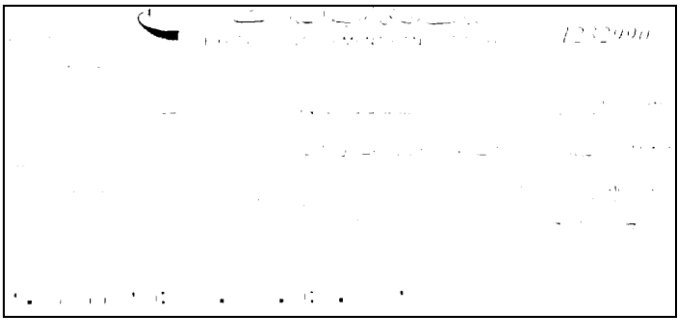


Figure c

9	10	11
6	7	8
3	4	5
0	1	2

Figure d

Figure1: (a) Thresholded check image; (b,c,d) Erosioned check image; (d) logo localization

III. The Classification Tree

Image classification is the problem of classifying images into known semantic classes. Let C = (C1, C2, …, Cn) be the check classes known a priori. We assume that we have a set T of training check images whose class membership is known and I of images that need to be classified. We want build a classification tree from training images. At each level of the classification tree, we aim to choose the best modeling of the training data based on the use of prototypes and the feature to determine the selected model.

The decision tree algorithm is a method for classifying objects based on a tree structure, where every internal node of the tree contains a test condition and every leaf node has an associated class label.

Decision tree is one of the simplest classifiers in machine learning. One of its advantages lies in its ease of interpretation as to which are the most distinguishing features in a classification problem [17]. In this section we investigate a three-level decision tree classification approach that first attempts to assign a query image to a set of classes within the database. The class rules were identified using the statistical features extracted from the colored histogram, the entropy and the energy based on training data. Each subclass may contain one or more type of checks. Thus, in the second level each check model in these subclasses is classified into two groups according to the existence or non-existence of logo. In the third level, check models that has logos are divided into several groups according to the location of the logo. Finally, if the selected class has only one model, the check type is detected. Otherwise, the logo or the bank name will be extracted from the processed image and it is correlated with the bank logos or bank names entry of model belongs to the same subclass. The check type that has a correlation exceeds a certain threshold and achieves the maximum value is therefore the investigated check model. Else, if all the obtained correlations are less than the threshold the check will be classified as unknown, and its model must be added to the check model database.

IV. Extraction of Check Fields

In order to recognize the various information fields, we should segment the image into the target object regions and extract them one by one. In this work we propose an algorithm to extract separately

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

filled-in items from Arabic bank-checks based on Hough transform. Hough transform is a method for estimating the parameters of a shape from its boundary points. Its principal concept is to define a mapping between an image space and a parameter space. The parameter space is defined by the parametric representation used to describe lines in the picture plane. Typically points or edges are mapped into a Hough space by a voting procedure. This voting procedure is carried out in the parameter space, from which object candidates are obtained as local maxima in an accumulator space that is explicitly constructed by the Hough transform algorithm [18].

The standard Hough Transform for straight line is represented by equation (1), where (x, y) denotes a point in the image space, and (r, θ) denotes its parameter in the polar coordinates, also called the Hough Transform parameter space. All points on the same line in the image space will intersect at one point in the Hough Transform parameter space.

$$r = x \cos\theta + y \sin\theta. \quad (6)$$

where r is the length of a normal from the origin to this line and θ is the orientation of r with respect to the X-axis (see figure 2). For any point (x,y) on this line, r and θ are constant.

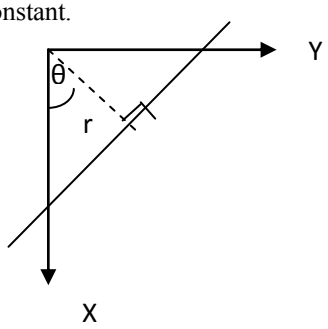


Figure 2. representation of a straight line in Houph parameters r, θ.

The extraction of segmented fields is accomplished through the following stages. In the first stage, the preprocessed bank check is converted into binary image. In this process a threshold of 128 is selected such that values greater than the threshold are mapped to white where as the values less than that are marked as black. Then, the erosion morphological operation is applied. Afterward, the eroded image is subtracted from the binary image in order to eliminate any noise and reduce the effect of stamps and handwritten information. Next, the Hough transform is applied on the resulted image in order to extract the printed lines. After the detection of all horizontal lines, those having a maximum value, within a certain threshold, of the accumulator are displayed. This threshold is chosen in away to allow us screening all guidelines in the processed image. However, due to check borders or back ground drawing some extracted lines may not be guidelines. Thus, these lines are ignored either automatically (by disregarding all extracted lines that lies near the four borders of the bank check) or by human intervention. Furthermore, some field may be missed due to missed guidelines in the original bank checks. Finally, the areas around the baselines are extracted.

Figure 3 represents an example of the developed field extraction process. It can be noted that some handwritten information (the signature) did not extracted correctly because it is outside the correct zone. This problem can be solved either by extracting the missed field manually or by instructing the customer to write in the correct zones.

V. Experimental results

This section describes the results of experiments performed with the algorithm described in the previous sections. The system introduced in this work proceeds in two phases: a training phase and a testing phase. The training phase has a set of filled check images and empty forms of the same check banks taken as reference. The test phase has a set of arbitrary check images.

The proposed system has been tested on various Arabic bank checks. The system introduced in this work proceeds in two phases: a training phase and a testing phase. All images are scanned off-line in a fixed frame size and are stored as 256 colored bitmap file. Then, a preprocessing algorithm [5] is accomplished. Afterward, we classify each image in the training data using two-step classifier. In the first step a three-level decision tree classifier is utilized, while in the second step. In the first level, statistical histogram features, entropy and energy are extracted from the enhanced colored check image. These features are used to classify the check images into several groups; each group may contain more then check types. Thus, in the second level, within each group the checks are categorized into two classes; with and without logo according to the with-logo and without-logo. Moreover, checks that have logos are classified into several classes according to the logo position. Finally, within the selected check class the check logo or check name is used to denote the check prototype.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

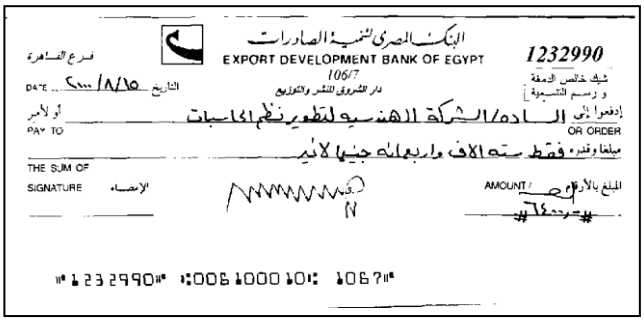


Figure 3-a

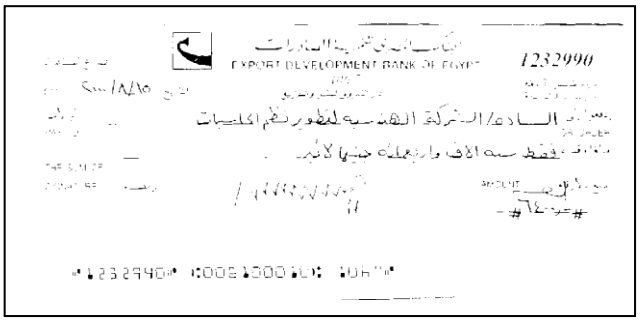


Figure 3-b

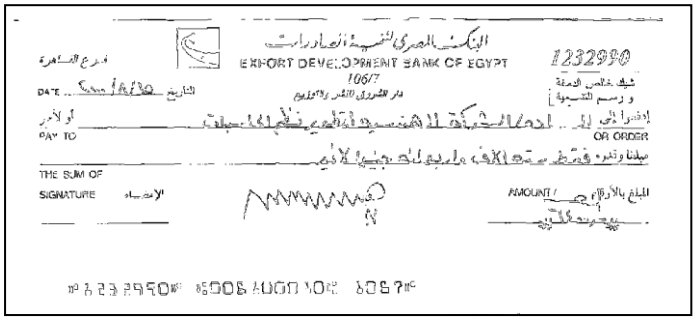


Figure 3-c

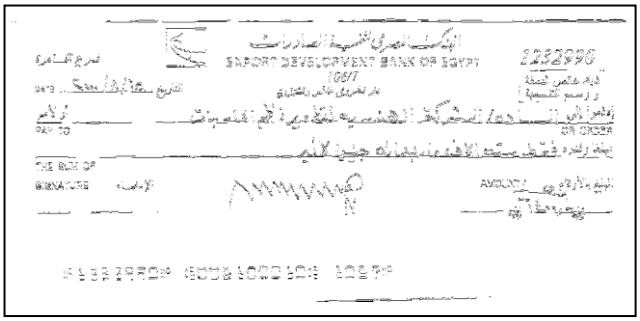


Figure 3-d

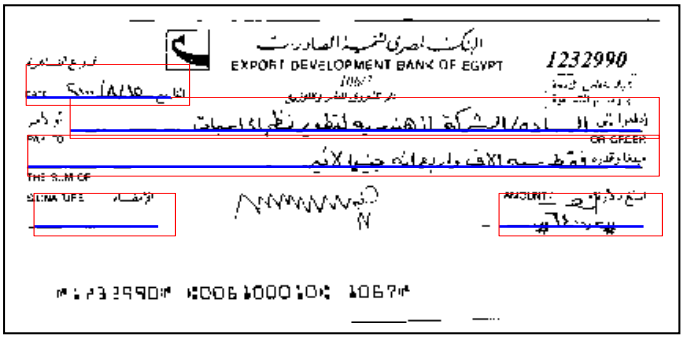


Figure 3-e

Figure 3: (a) Binary check image; (b) the eroded image; (c)the eroded image subtracted from the binary image; (d) applying Sobel filter; (e) the extracted fields .

The average execution time on the selected images is about 12.5 sec per one input image, matched with 24 different kinds of bank checks. This time is comparable to the algorithm presented in [6] but this algorithm saves the time of training the genetic algorithm. Furthermore, the proposed feature extraction process has less time and space complexity than that proposed in [6].

The system was trained on 24 different Egyptian bank checks, while 40 check images were used to test the proposed system. Owing to privacy and confidentiality laws, there are no publicly available standard or benchmark check databases to apply different techniques or to perform a comparative analysis. Hence, we implement our system on the database mentioned in [7]. The images were scanned with 300 dpi and 256 colors. The results show that the approach is effective and the proposed algorithm performs correctly on all the test checks.

Furthermore, Houph transform is utilized to detect guidelines that in turn used to extract the handwritten fields. The extraction rate of this method is about

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

97.33% for all handwritten components. The error in the extraction rate is caused by the error in extraction of the horizontal lines as well as that customers may write their information outside the proper zones.

VI. CONCLUSION

This paper proposes a new technique for check classification as well as handwritten field extraction from bank checks. The proposed technique is based on a two-level classification approach. The first-level attempts to assign the input image to a restricted set of classes within the data base using decision tree classification, while the second-level returns the best matches to each of the selected classes. The results indicate that the proposed approach is effective and its performance is encouraging. On the other hand, an algorithm based on Houph transform is applied to extract the user-entered data (numerical amount, literal amount and date). The results obtained demonstrate that the implemented algorithm is effective and that some errors can be minimized by influencing the customers to write in the correct zones.

REFERENCES

3. Palacios, R., Gupta, A.: A system for processing handwritten bank checks automatically. Image Vis. Comput. 26(10), 1297-1313 (2008). J. Blodgett, "Beyond cheque image statements: A new strategy for the 1990s", Advance Imaging, Vol. 9, pp. 73-75, 1994.
4. J. Blodgett, "Beyond cheque image statements: A new strategy for the 1990s", Advance Imaging, Vol. 9, pp. 73-75, 1994.
5. Automatic processing of handwritten bank cheque images: a survey, R. Jayadevan, S. R. Kolhe, P. M. Patil and U. Pal, International Journal on Document Analysis and Recognition,SpringerLink,15 July 2011
6. C.Y. Suen, Q. Xu and L. Lam, "Automatic recognition of handwritten data on cheques – Fact or Fiction?", Pattern Recognition Letters, Vol.20, pp. 1287-1295, 1999.]
7. Heba A. Elnemr, Mohsen Rashwan, Ahmed Hussien, Mohammed S. Elsherif, "Automatic Classification of Bank Checks," Poroceedings of IEEE ISSPIT2001, Cairo, Egypt, Dec. 2001.
8. Elnemr, H.A.; Rashwan, M.; Elsherif, M.S.; Hussien, A., "Hierarchical Classification of Bank Checks Using Genetic Algorithms", The 3rd International Symposium on Image and Signal Processing and Analysis ISPA 2003 September 18-20, 2003, Rome, Italy, pp. 770-773, Vol.2.
9. Heba A. Elnemr, "Automatic Processing of Bank Checks", M.Sc. Thesis, Cairo University, 2003.
10. Mohit Mehta, Rupesh Sanchati and Ajay MarchyaInternational, "Automatic Cheque Processing System", Journal of Computer and Electrical Engineering, Vol. 2, No. 4, August, 2010, pp. 761- 765.
11. Vamsi Krishna Madasu, Brian Charles Lovell, "Automatic Segmentation and Recognition of Bank Cheque Fields", Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA 2005), 2005, ISBN: 0-7695-2467-2, pp. 33
12. Md. Rezaul Hoque Khan, Gahangir Hossain , Bengali Handwritten Bank check Recognition Using Automatic Extraction of the User–Entered Data, proceedings of 8 th ICCIT pp. 650-654, 2005].
13. Palacious, R, Gupta, A.: A System for Processing Handwritten Bank Checks Automatically, Image and vision Computer , volume 26, issue (10), pp. 1297-1313, October,2008
14. Fadoua BOUAFIF SAMOUD1, Samia Snoussi Maddouri, Haikal El Abed, Noureddine Ellouze, Comparison Of Two Handwritten Arabic Zones Extraction Methods Of Complex Documents, Proceedings of International Arab Conference on Information Technology, pp. 1-7, 2008
15. Sofiene Haboubi, Samia S. M., Extraction of handwritten areas from colored image bank checks by an hybrid method, International Conference on Machine Intelligence (ACIDCA-ICIM), Tozeur, Tunisia, November 2005.
16. Liu, K., Suen, C.Y., Cheriet, M., Said, J.N., Nadal, C., Tang, Y.Y.: Automatic extraction of baselines and data from check images. Int. J. Pattern Recognition Artificial. Intelligence. volume 11, issue (4), 675-697(1997)
17. Ke Liu , Ching Y. Suen , Christine Nadal Automatic Extraction of Items from Cheque Images for Payment Recognition, Int. Proceedings of 13 th ICPR, pp. 798-802, 1996
18. Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing", Addison-Wesley, June, 1992.
19. D. LU and Q. WENG, "A survey of image classification methods and techniques for improving classification performance", Vol. 28, No. 5, 10 March 2007, 823–870
20. R. Duda and P. Hart. "Use of the hough transform to detect lines and curves in pictures". Communications of the ACM, pp11–15, 1972.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

DATA AGGREGATION WITH ENERGY EFFICIENT RELIABLE ROUTING PROTOCOL FOR WIRELESS SENSOR NETWORKS

Basavaraj S. Mathapati
Dept. of Computer Science & Engg.
Appa IET, Gulbarga
Karanataka, India
b_ math@rediffmail.com

Siddarama. R. Patil
Dept. of Electronics & Comm. Engg.
P. D. A College of Engineering
Gulbarga, Karanataka, India
pdapatil@gmail.com

V. D. Mytri
Principal
GND College of Engineering, Bidar
Karanataka, India
vithalmytri@gmail.com

***Abstract* - In Wireless Sensor Networks (WSN), data aggregation is essentially used to gather and aggregate data in an energy efficient manner so that network lifetime is enhanced. Data aggregation protocols aims at eliminating redundant data transmission. Power consumption is an important aspect to be considered in the data aggregation which is a scarce resource and they are irreplaceable. In addition to power consumption, reliability is also of major concern in data aggregation. In this paper, we propose to design an energy efficient reliable data aggregation technique for wireless sensor networks. Initially we form clusters and a coordinator node (CN) is selected near the cluster in order to monitor the nodes in the cluster. The CN selects a cluster head (CH) in each cluster based upon the energy level and the distance to the CN. The packets sent by the sensor nodes are aggregated at the CH and transmitted to the CN. The CN measures the loss ratio and compares it with a threshold value of loss ratio. Depending upon this value, the forward node count is incremented or decremented and the cluster size is adaptively changed, ensuring reliability and balanced energy consumption. From our simulation results we prove that this technique is efficient in energy consumption and reliability.**

I. INTRODUCTION

In data aggregation, the aggregation processes are used to aggregate the sensor data effectively. Data aggregation techniques enhances the network lifetime by gathering and aggregating the data in an energy efficient manner. A striking method for data gathering in wireless sensor networks involves distributed system architectures and dynamic access via wireless connectivity. In the case of energy constraint wireless sensor networks, the data aggregation techniques intend to eradicate the redundant data transmissions thereby improving the lifetime of the network. [1]

Due to that the sensor nodes are tightly packed in the sensor networks, there are possibilities for the nearby sensor nodes to overlap sensing ranges. Because of this, redundant or correlated data are collected by the sensor networks. In order to save the energy, the data correlation is subjugated which effectively reduces the amount of data transmitted in the network. In wireless sensor network routing, data aggregation proves to be an important aspect. The data originating from different sensor nodes aggregate together in the sink node during transmission. [2]

The main purpose of the data gathering in wireless sensor network (WSNs) is to obtain valuable information from the operating environment. It has been proven that the data redundancy can be eradicated and the communication load can be reduced using the data aggregation techniques. Multiple data sources and a data sink are included in the typical communication patterns of data aggregation. A data aggregation tree is constructed using the transmitted packet and this is similar to the reverse multicast structure. [3]

Advantage of Data Aggregation: Robustness and accuracy of information acquired from the network can be improved effectively. The data aggregation requires the data fusion processing in order to reduce the redundant information which is present in the data collected from the sensor nodes. Traffic load is minimized and the energy in the sensors can be conserved with the help of data aggregation. Disadvantage of Data Aggregation: The cluster heads are also known as the data aggregator nodes which combine the data in order to send it to the base station. There are chances of malicious attackers in the cluster head or the aggregator node. The accuracy of the aggregate data sent to the base station cannot be guaranteed when the cluster head is compromised. The uncompromised nodes send several copies of the aggregate result to the base station which increases the power consumed at these nodes. [4].

In this paper, we proposed to develop A Data Aggregation with Energy Efficient Reliable Routing Protocol for Wireless Sensor Networks (DAEERRP) which is energy efficient and reliable. This technique is based on cluster formation and the loss ratios of the clusters are measured so that the energy consumption can be effectively reduced. Reliable transmission can be provided in the clusters using a coordinate node.

The reminder of this paper is organized as follows. In Section II, we introduce previous work related to our study. Section III gives description of proposed protocol. The simulation results in Section IV and in Section V paper is concluded.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

II. RELATED WORK

Ren P. Liu et al [5] have proposed an Efficient Reliable Data Collection (eRDC) algorithm. Maximum number of retransmissions is controlled in order to achieve energy savings. Dynamic programming concept is used to find the optimal solution. Implementation of eRDC is provided which uses next hop link quality and number of hops for determining number of retransmissions.

Volker Turau et al [6] have presented the design and preliminary evaluation of a reliable data gathering service of periodic data in the face of poor link quality and frequent disconnects. The data is buffered by persistent storage provided by the nodes using services based on a packet-level, and hop-by-hop routing protocol. This design also provides an upper limit for sampling rate that is handled reliably.

Hemant Sethi et al [7] have proposed an Energy Efficient Interest Based Reliable Data Aggregation (EIRDA) Protocol for WSNs. Here each cluster considers the uniform distribution of sensor nodes using EIRDA which is a static clustering scheme. Beta-distribution function is used to provide reliability with the help of Functional Reputation concept. The overall impact of all measures taken at each phase of protocol implementation is clearly visible on the energy spent in the setup phase of the protocol.

David Gugelmann et al [8] have presented a novel data dissemination protocol with a focus on reliability and energy-efficiency. Scheduling image dissemination only during reserved time slots eliminates interference with the regular data gathering protocol and increases the observability during the network reprogramming phase.

Mingming Lu et al [9] have proposed the data-gathering problem in wireless sensor networks from the maximization of the expected network utility point of view. The resource scarcity and the unstable nature of wireless channels are considered here. Data gathering problem is designed as an optimization problem and the NP-hard problem is proved here. For both the broadcast tree and reverse multicast tree problems several heuristics were proposed.

Numerous sensor nodes are deployed densely in WSN. The sensor nodes forward the packets generated during the event detection to the base station. The neighboring nodes help in the transmission of packets to the sink. These packets are processed after they reach the base station. Thus reliable event detection requires reliable data transport in the WSNs. The reliable data deliveries in wired networking are not apt for TCP/IP protocols due to its inaccurate nature. [10]
In wireless sensor networks, the data aggregation needs to be processed based upon the energy and the reliability due to the following reasons:

- Operate the energy resources over a long period.
- Efficient energy is required for the sensor nodes in order to schedule their transmission strictly. Improper transmissions may lead to idle listening and overhearing causing energy wastage.
- The aggregation increases the amount of data concentrated in a single message which requires alteration in reliability.
- In event detection, the packets are transmitted from the sensor nodes to the base station and then to the neighboring which is possible through reliable data transport.

To overcome our previous certain issues on data aggression, we provide an efficient energy based and reliable data aggregation technique in the wireless sensor network. This technique is based on cluster formation and the loss ratios of the clusters are measured so that the energy consumption can be effectively reduced. Reliable transmission can be provided in the clusters using a coordinate node.

III. DESIGN OF DAEERRP

A. Overview

In this paper, we propose to develop a data aggregation technique which is energy efficient and reliable. Initially a cluster is formed and the cluster head is selected based upon the cost value which is explained in section 3.2. The nodes in the cluster maintain a Neighbor information table (NIT) containing Node id, Distance and Cost. This NIT information is sent to the cluster head. Each cluster selects a coordinator node (CN) randomly in the network which is closer to the cluster and monitors the operations of the sensor nodes and commands them for specific operations. The cluster head aggregates the data and sends it to the CN.

The CN calculates the loss ratio which is the ratio of number of packets dropped and total packets broadcast from the source. Based upon the loss ratio, the cluster size can be modified and the forward node count of each node can be incremented or decremented which is explained in section 3.4. Once the cluster size is changed, the CN gathers the information again from the cluster head compresses it and sends it to the sink.

Since the loss ratio is measured at the CN itself, the energy consumption can be effectively reduced. Also the reliability can be increased due to altering the cluster size before the data is transmitted to the sink.

B. Cluster Head Selection

- Initially the sensor nodes are arranged into clusters and the CN selects the cluster head for each cluster.
- Numbers of neighboring nodes M are determined by the CN based upon the node density.
- The sensor nodes transmit the M number of nearest neighbors to the CN.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Received signal strength indicator (RSSI) estimates the distance to the nodes.
- K-theorem is used in each cluster by the CN to select the candidate set of cluster heads (S_{CH}).
- The request for the candidate set of cluster heads is sent by the CN and the sensor nodes reply their cost value (CV).
- Each candidate cluster head node calculate its own CV based on residual energy, and distance to coordinator node, and send it to CN. Calculation of cost value is explained in section 3.2.2.
- The coordinator node selects a node as cluster head among candidate set of cluster heads for each cluster based on CV. The higher the CV a node has; greater the chances of being cluster head. The CN confirms each cluster about their CH.

In this figure 1, we assume a cluster with three sensor nodes. The CN sends a request for the nodes in the cluster. The cost value of the sensor nodes 1, 2, and 3 are sent back to the CN. The node having higher CN becomes the Cluster head. Here the node 1 has CV value as 7 and it is elected as the CH for the cluster and the CN sends this information to other Clusters.

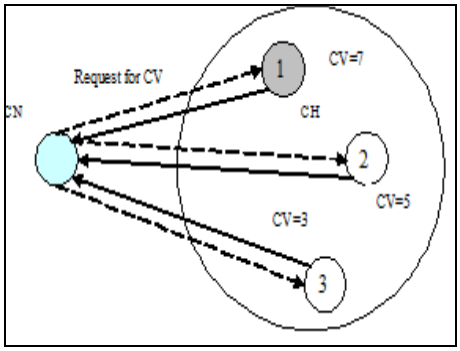


Figure 1: Selection of Cluster Head.

Selection of M for clustering:

In each cluster, value of M is set by the CN and this M value is relative to the node density and the ratio of cluster heads in a WSN.

The ratio should be below 0.50 ranging from 0.01 to 0.99. Many local optima can be obtained when the value of M is lesser. The number of best sensor nodes which are suitable for CH can be determined using M value. Optimal sensor node for cluster head can be selected by providing alternate suboptimal options by M.

M nearest neighbors is selected for each sensor node deployed in the cluster which is based on the distance. The received signal strength indicator calculates the distance between the sensor nodes. For larger distance, nearest neighboring nodes can be determined using multihop communication route. Energy consumption is less in choosing a neighbor in multi-hop connection when compared to the direct communication.

Every sensor node calculates its frequency of occurrence and minimum frequency required for a cluster to become a CH is also calculated. The weighted mean of frequencies is calculated and it is enhanced by adding 1 to it.
The product of each frequency of occurrence and number of sensor nodes having that frequency is calculated as weighted mean. The frequency value is rounded to its nearest integer. The value of Sensor nodes having frequency F or greater are identified and they become the candidates for cluster head (CH). The candidate cluster head nodes would always be equal to value of F [10].

Cost Value Calculation:

The Cost value (CV) is calculated based on following criterion:

Residual energy (E) : The residual energy of a node preferably is greater than the approximate energy dissipated in previous round by the cluster head.

Distance to coordinator node (D) :The nodes having less distance from coordinator node should have higher probability to become cluster head. As energy consumption is directly proportional to the square of distance.

Cost value is based on the residual energy and the distance to the coordinate node. The cost is high, when the residual energy is high and the distance to the coordinator node is less.

$$CV = (a \times E) + (b \times \frac{1}{D}) \quad \ldots. (1)$$

where *a* and *b* are normalization constants.

C. Loss Ratio Calculations

Each node maintains a forward node count (C_{FN}), which denotes the broadcast or rebroadcast probability.
Initially C_{FN} [Nk] = C_{FN}min, for all nodes Nk, k=1, 2……. C_{FN}min is the minimum number of forwarding nodes. Without loss of generality, we can assume that
C_{FN}min =1. The steps involved in the adaptive energy efficient forwarding phase are given below:

- Suppose N wants to send the collected data to the sink, it attaches its cost to the data packet and broadcast the packet to the nearest neighbors.
- When a neighbor N1 receives the packet from N, it first checks its cost is less than that of N. If it is less, it further forwards the packet. Otherwise it drops the packet, since N1 is not towards the direction of the sink.
- When the packet reaches the destination D, it measures the loss ratio (LR), which is the ratio of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

number of packets dropped and total packets broadcast from the source.

- Then D sends this LR value as a feed back to the source N.

When N receives this value, it checks the value of LR. It then modifies the value of C_{FN} as

$$C_{FN} = C_{FN} + \gamma, \text{ if } LR > LRmax. \quad \ldots. (2)$$

Where γ is the minimum increment of decrement count and LRmax is the maximum threshold value of loss rate.

It then rebroadcast the data packets with the incremented C_{FN}, so that increasing the reachability of the sink. The total power required to reach the sink is thus calculated based on the cost field of all the nodes in C_{FN}. For example, if C_{FN} =4, then the minimum required power will be 4 * cost of each neighbor node in the NIT.

When the rebroadcast packets reach the destination D, it again calculates the loss ratio LR and sends back to N. It then reassigns the value of C_{FN}, depending on the value of LR. Once LR < LRmax , then

$$C_{FN} = C_{FN} - \gamma, \text{ until } C_{FN} >= C_{FN}min \quad \ldots.. (3) [6]$$

This data aggregation technique proves to be efficient in terms of energy and reliability since,

- Delay can be reduced due to that the loss ratio is measured in the CN itself. Measuring loss ratio at the sink causes high delay.
- Energy is reduced effectively when the size of the cluster is altered based upon the loss ratio.

The reliability can be maintained due to the change in the size of the clusters.

IV. Simulation Results

A. Simulation Setup

We evaluate our DAEERRP scheme through NS2 simulation [11]. We considered a random network deployed in an area of 500 X 500 m. The number of nodes is varied as 100,125,150,175 and 200. Initially the nodes are placed randomly in the specified area. The sink is assumed to be situated 100 meters away from the above specified area. The initial energy of all the nodes assumed as 8.1 joules. In our simulation, the channel capacity of mobile hosts is set to the same value: 2 Mbps. We use the distributed coordination function (DCF) of IEEE 802.11 for wireless LANs as the MAC layer protocol. The simulated traffic is CBR with UDP source and sink. All experimental results presented in this section are averages of five runs on different randomly chosen scenarios. The Table- I summarize the simulation parameters used.

B. Performance Metrics

We compare DAEERRP with the extended DPC-EERP [12] scheme. We evaluate mainly the performance according to the following metrics.

- **Average end-to-end delay:** The end-to-end-delay is averaged over all surviving data packets from the sources to the destinations.

- **Average Packet Delivery Ratio:** It is the ratio of the number .of packets received successfully and the total number of packets transmitted.

No. of Nodes	100,125,150,175 and 200
Area Size	500 X 500
Mac	802.11
Simulation Time	50 sec
Traffic Source	CBR
Packet Size	512
Transmit Power	0.360 w
Receiving Power	0.395 w
Idle Power	0.335 w
Initial Energy	8.1 J
Transmission Range	75m
Rate	50,100,150 and 200Kb.

TABLE I SIMULATION PARAMETERS

- **Drop:** It is the total number of packets dropped during the transmission.

- **Energy Consumption**: It is the average energy consumption of all nodes in sending, receiving and forward operations.

The simulation results are presented in the next section.

4.3. Simulation Results

In our first experiment we vary the number of nodes as 100,125,150,175 and 200.

The figure 2, we can see that the average end-to-end delay of our proposed DAEERRP protocol is less than the existing DPC-EERP protocol.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

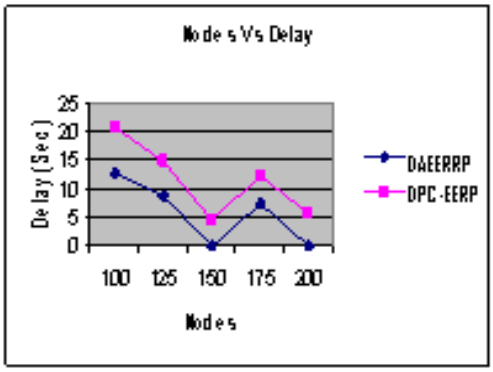


Fig 2: Nodes Vs Delay

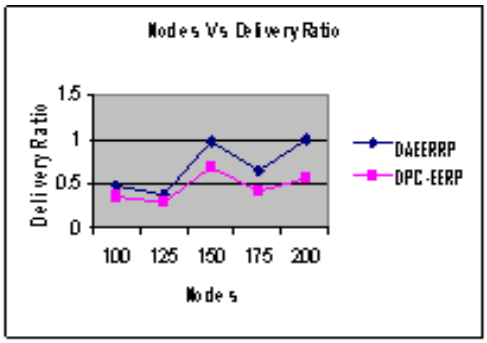


Fig 3: Nodes Vs Delivery Ratio

From the figure 3, we can see that the Delivery Ratio of our proposed DAEERRP is higher than the existing DPC-EERP protocol.

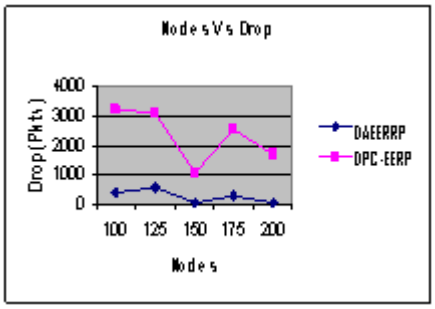


Fig 4: Nodes Vs Drop

From figure 4, we can see that the packet drop of our proposed DAEERRP is lower than the existing DPC-EERP protocol.

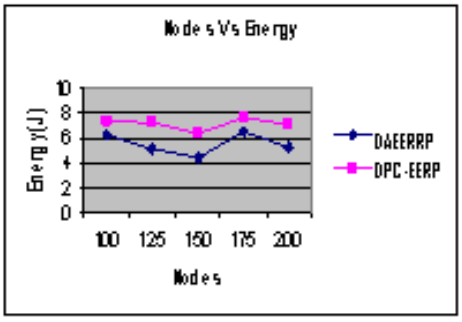


Fig 5: Nodes Vs Energy

From figure 5, we can see that the Energy consumption of proposed DAEERRP is less than the existing DPC-EERP protocol.

V. CONCLUSION

In this paper, we have provided efficient data aggregation technique which considers both energy and reliability. Initially, the network is partitioned to various clusters .Each cluster selects a coordinator node (CN) randomly in the network which is closer to the cluster and monitors the operations of the sensor nodes and commands them for specific operations. In each cluster, the cluster head is selected based upon the cost value. The nodes in the cluster maintain a Neighbor information table (NIT) containing Node id, Distance and Cost. This NIT information is sent to the cluster head.. The cluster head aggregates the data and sends it to the CN. The CN calculates the loss ratio which is the ratio of number of packets dropped and total packets broadcast from the source. Based upon the loss ratio, the cluster size can be modified and the forward node count of each node can be incremented or decremented. Once the cluster size is changed, the CN gathers the information again from the cluster head compresses it and sends it to the sink. This technique proves to be efficient since the delay is reduced due to that the loss ratio is measured in the CN itself. Energy is reduced effectively and reliability is maintained when the size of the cluster is altered based upon the loss ratio. From our simulation results we prove that this technique is efficient in energy consumption and reliability.

REFERENCES

[1] Nandini. S. Patil, and Prof. P. R. Patil "Data Aggregation in Wireless Sensor Network", IEEE International Conference on Computational Intelligence and Computing Research, 2010.

[2] Kevin Yuen, Baochun Li, and Ben Liang "Distributed Minimum Energy Data Gathering and Aggregation in Sensor Networks" Communications, IEEE International Conference, 2006. ICC '06.

[3] Yang Yu, Bhaskar Krishnamachari, and Viktor K. Prasanna "Energy-Latency Tradeoffs for Data Gathering in Wireless Sensor Networks" , . Twenty-third Annual joint Conferences of the IEEE Computer and Communications Societies, INFOCOM 2004.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[4] Kiran Maraiya, Kamal Kant, and Nitin Gupta "Wireless Sensor Network: A Review on Data Aggregation" International Journal of Scientific & Engineering Research Volume 2, Issue 4, April -2011.

[5] Ren P. Liu, John Zic, Iain B. Collings, Alex Y. Dong, and Sanjay Jha "Efficient Reliable Data Collection in Wireless Sensor Networks", IEEE 68 conference Vehicular Technology Conference, VTC 2008. .

[6] Volker Turau and Christoph Weyer "Long-term Reliable Data Gathering Using Wireless Sensor Networks" IEEE 4 th international Conference on Networked Sensing Systems, 2007. INSS ' 07. .

[7] Hemant Sethi, Devendra Prasad, and R. B. Patel "EIRDA: An Energy Efficient Interest based Reliable Data Aggregation Protocol for Wireless Sensor Networks" International Journal of Computer Applications (0975 – 8887) Volume 22– No.7, May 2011.

[8] David Gugelmann, Philipp Sommer, and Roger Wattenhofer "Poster Abstract: Reliable and Energy-Efficient Bulk-Data Dissemination in Wireless Sensor Networks" SenSys'10, November 3–5, 2010.

[9] Mingming Lu and Jie Wu "Utility-Based Data-Gathering in Wireless Sensor Networks with Unstable Links", ICDCN Proceedings of the 9th international conference on 2008.

[10] Muhammad Imran, Asfandyar khan, and Azween B . Abdullah "Energy Balancing Through Cluster Head Selection Using K-Theorem In Homogeneous Wireless Sensor Networks" International Conference on Science & Technology: Applications in Industry & Education (2008).

[11] Network Simulator: http://www.isi.edu.nsnam/ns.

[12] Basavaraj.S.Mathapati,,Siddaram.R.pati,V.D.Mytri,"Distributed power control with Energy Efficient Routing Protocol for wireless Sensor Network", International Journal of Information Technology and knowledge Management ,2012.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The requirements of Parallel Data Warehousing Environment to Improve the Performance with dominating sets for Next generation Users

Umapavankumar Kethavarapu
Research Scholar at Pondicherry Engineering College,
CSE Dept, Pondicherry
Pondicherry, India
umapawanmtech@gmail.com

Dr.S.Saraswathi
Associate Professor IT DEPT
Pondicherry Engineering College,
Pondicherry, India
swathimuk@yahoo.com

***Abstract*—The data warehousing (DWH) environment is useful to handle bulk data processing by providing techno-functional aspects, the parallelism and distribution in DWH will greatly serve the customers and management. In this paper we described the master data management (MDM), data virtualization, and integration of middle ware technologies to the parallel data warehouse (PDW).It is very much useful because the users are requiring their data in less time and they want to interact with the system in less complex way. The management of the DWH also wanted to handle their data in efficient manner so as to produce strategic decisions. To achieve all the requirements management of the Master data, data virtualization and as well as middleware technologies (MWT) usage in the PDW giving a better solution to both developers and customers. The proposed work gives a betterment in extraction, Transformation and Loading(ETL)process, management of common data in data marts, data warehouse and middleware benefits to the PDW.To handle faster data processing and work distribution in PDW ,MDM,data virtualization and MWT we proposed one common solution is usage of Dominating sets in the identification of systems and transactions which are participating in data processing the domain of participated systems we named as critical sub-system and the transactions activated are known as critical transactions.**

Keywords-Data warehousing; parallel Data warehousing; Master Data Management; Data Virtualization; Middle ware technologies; dominating sets; Critical Transactions.

I. INTRODUCTION

Dominating set in a graph G = (V, E) is a subset D of V such that every node u € V is in D or adjacent to some node v€ D. A dominating set D is called Connected Dominating Set (CDS) if it is a induced connected sub graph of G. A Minimum Connected Dominating Set (MCDS) is a connected dominating set with smallest possible cardinality. Among all the CDSs of G. Connected Dominating Sets are popularly used for constructing virtual backbones for broadcasting operation in data passing [1]. Establishing a virtual data passing path in data movement is an important issue because it reduces unnecessary message transmission or flooding in the network. It helps in reducing unnecessary systems in the data processing because a limited number of sensors are engaged in message transmission and thus it helps in improving the data quality in real-time data processing in the DWH, PDW. Data virtualization requires where to replicate the data according to request made by the users. The same concept is also helps to improve performance considerations by identifying critical sub-systems and critical-transactions in case of MDM and MWT in DWH environment. So the dominating sets usage is a common solution to quality data processing in case of DWH, PDW, MDM, Data Virtualization and MWT.

The DWH environment is a better approach to gather the data from various locations and from various sources in various formats. The DWH provides a repository of data which includes processing of the data from various online transaction processing sources and will give a common interface to access and to process the data to achieve strategic decisions The details about DWH environment is provided in section [2] The concept of parallelism is very much needed in the maintenance of bulk data and various categories of users ,as DWH consists of bulk data and various categories of the users in the form of functional and technical aspects the integration of parallel aspects to DWH will yields better results. The description about PDW is specified in section [3].The real-time data management in DWH is much complex if we are using incremental approach to populate the data .The main concern of the DWH is to maintain the data quality. Real-time data quality is not that much of easy because the data is available from different sources and in different formats. The initial data captured by the organization will be continued in all the ETL and reporting purpose, the initial data which is commonly used for the purpose of ETL and reporting is known as Master data. MDM is very much important in case of normal DWH environment and as well in case of PDW environment. We described the management of the MDM in section [4] which will gives the details about the MDM in normal DWH and PDW. Data virtualization is one more important factor in the efficient data management in DWH and as well PDW. Data visualization specifies availability of the data to various locations with the property of transparency which will improve performance of access and process of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

queries. The detailed description about data virtualization is available in section [5]. The MWT work is very good ,they provides user friendly environments to any kind of complex applications and user can make use of his/her own technique of accessing the data and processing the data. Section [6] describes the way of integration of MWT with DWH and PDW so as to strengthen the usage of data warehouse by various categories of the users with their own method of implementation irrespective of tool knowledge of DWH. The overall aim of the paper is improvement of the data processing in DWH and PDW environment by maintaining quality in MDM, Data virtualization is also used for the quality improvement of the DWH and PDW in the query processing speed application response and availability of the data to the users located geographically. The another technique we proposed in this paper is embedding the MWT such as remote procedure call(RPC),Common Object Request Broker Architecture(CORBA) and open systems such as C,C++ etc usage in the handling of DWH and PDW. We are sure that by achieving all these aspects with respect to DWH and PDW we can expect betterment in the processing speed and availability of the data in the DWH environments. With the integration of MWT with DWH we can expect more user interaction and faster processing. If implementation is in Open sources such C-language or CPP then the usage is better because the developers are familiar with these open source systems and the maintenance cost is also less compared with other middle ware technologies with server configurations

II. Data Warehousing Environment

A data warehouse is a collection of data created to support decision making applications. Improve decision making.

Support key corporate initiatives such as performance management, B2C and B2B e-commerce, and customer relationship management the characteristics of data warehouse are Subject-oriented, Time-variant, Integrated, Non-volatile. Data warehousing is the entire processing of Extraction, Transformation and Loading of data to the data warehouse and access of the data by end users and applications. A data mart stores data for a limited number of subject areas. It is used to support specific applications. An independent data mart can be created directly from the source systems. A dependent data mart is populated from the data warehouse. An operational data store consolidates data from multiple source systems and provides a near real-time, integrated view of volatile, current data. Its purpose is to provide integrated data for operational purposes. It has add, change, and delete functionality. Two data warehousing strategies are available. First is Enterprise-wide warehouse, top down, the Inmon methodology. Second one is Data mart, bottom up, the Kimball methodology. Data are moved from source to target. Meta Data Integration, A growing realization that meta data is critical to data warehousing success. Vendors like Microsoft, Computer Associates, and Oracle have entered the meta data marketplace with significant product offerings. On-line Analytical Processing (OLAP), A set of functionality that facilitates multidimensional analysis. Star Schema Creates non-normalized data structures [2].

A data warehouse is a collection of data created to support decision making applications. Improve decision making.

Support key corporate initiatives such as performance management, B2C and B2B e-commerce, and customer relationship management the characteristics of data warehouse are Subject-oriented, Time-variant, Integrated, Non-volatile. Data warehousing is the entire processing of Extraction, Transformation and Loading of data to the data warehouse and access of the data by end users and applications. A data mart stores data for a limited number of subject areas. It is used to support specific applications. An independent data mart can be created directly from the source systems. A dependent data mart is populated from the data warehouse. An operational data store consolidates data from multiple source systems and provides a near real-time, integrated view of volatile, current data. Its purpose is to provide integrated data for operational purposes. It has add, change, and delete functionality. Two data warehousing strategies are available. First is Enterprise-wide warehouse, top down, the Inmon methodology. Second one is Data mart, bottom up, the Kimball methodology. Data are moved from source to target. Meta Data Integration, A growing realization that meta data is critical to data warehousing success. Vendors like Microsoft, Computer Associates, and Oracle have entered the meta data marketplace with significant product offerings. On-line Analytical Processing (OLAP), A set of functionality that facilitates multidimensional analysis. Star Schema Creates non-normalized data structures [2].

Figure 1. Three tiers of BI environments

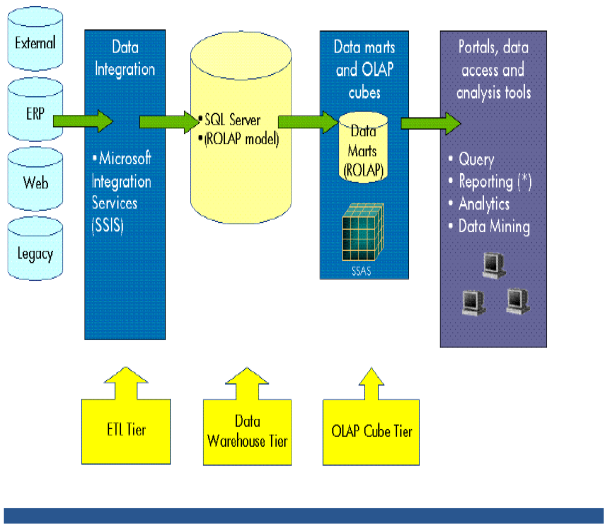


Fig. 1 3-tier DWH

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. PALLELISM IN DATA WAREHOUSING

A Parallel processing is another method used to improve performance in a computer system, when a system processes two different instructions simultaneously, it is performing parallel processing [3]. Saves time, Cost savings, overcoming memory constraints, Granularity – The ratio of computation to communication. Coarse – High computation, low communication, Fine – Low computation, high communication. Parallel Programming Models, Shared Memory, Threads, Messaging Passing, Data Parallel. Shared Memory Model Appears to the user as a single shared memory, despite hardware implementations. Program development can be simplified since there is no need to explicitly specify communication between tasks. Thread model a single process may have multiple, concurrent execution paths. A single process may have multiple, concurrent execution paths. Programmer is responsible for determining all parallelism. Message Passing Model, Tasks exchange data by sending and receiving messages. Typically used with distributed memory architectures. Data Parallel Model Tasks performing the same operations on a set of data. Each task working on a separate piece of the set. Works well with either shared memory or distributed memory architectures. As DWH having storage and processing of the data to generate strategic decisions in the form of reports we suggested that usage of threads and data parallel are better approaches to effectively implementing the parallelism for DWH environments. In DWH granularity specifies up to what extent the data is processed, for example the DWH granularity is observed the hierarchy as Schema->Data Base->Table->Record->Rows->Columns->Value. According to parallel aspects two granularities are observed, Coarse and Fine grained in coarse the communication is less and computation is high, which is reverse in case of Fine grained. But in DWH there is exists communication between the servers and users and computation is also required, so we proposed a federated model for DWH environment to achieve parallelism and we name it as hybrid grained federated DWH.Due to this approach we can have both communication and computation in mixed mode. The immediate benefits of parallelism to DWH are Speed-Up and Scale-Up.

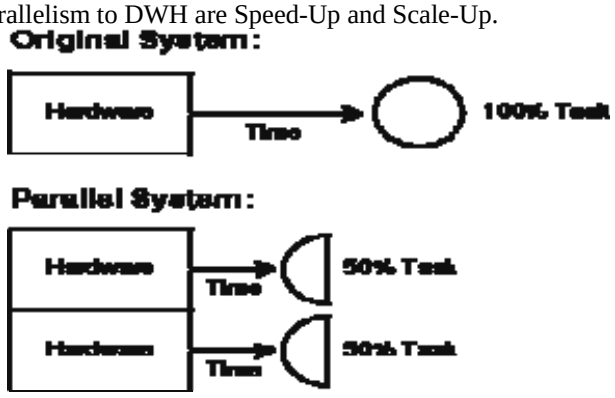


Fig. 1 speed up in parallel systems

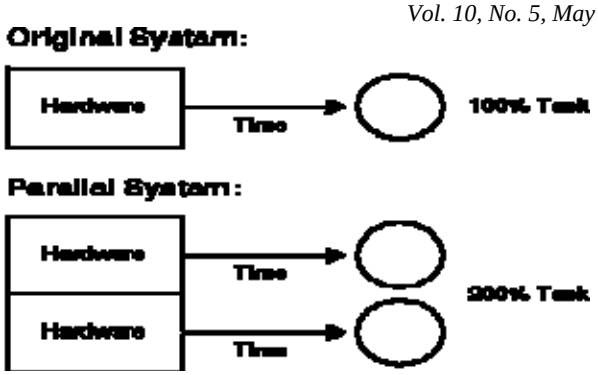


Fig. 3 scale- up in parallel systems

IV. CASE STUDY OF PDW

Organizations will begin to acquire hundreds of terabytes if not petabytes of data. While symmetric multi-processing (SMP) works well for small data warehouses and data marts, many organizations are adopting massively parallel processing (MPP) architectures to effectively manage, store, and unlock valuable business insights from complex data. Microsoft® SQL Server® 2008 R2 Parallel Data Warehouse is built on MPP technology that provides enterprise-class performance and scalability, flexibility and choice of hardware vendors, and the most comprehensive data warehouse solution through integration with both Microsoft and non-Microsoft Business Intelligence tools [4]. Parallel Data Warehouse offers flexibility and choice with leading hardware vendors such as HP and Dell. Parallel Data Warehouse offers more than just a data warehouse engine. It offers a comprehensive data warehouse solution with a complementary set of tools for ETL, analysis, reporting, MDM, and real-time data warehousing. The following diagram refers to the reference architecture of PDW provided by Microsoft Corporation [11]. Which will give an overview of how a PDW can provide incredible benefits to the functional and technical users and we can identify these benefits only in parallel environment which is not possible to achieve in normal data warehousing.

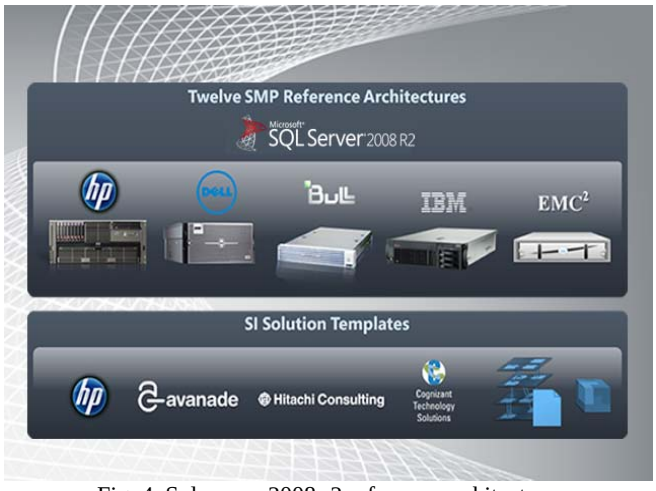


Fig. 4 Sql server 2008 r2 reference architecture

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The key benefits provided by PDW architecture to the users are

Key Features

Feature	Benefits
Cost based Query Optimizer and Enhanced Data Movement Services	• Breakthrough performance with up to 10X improvement • Predictable query performance
Integration with System Center Operations Manager	• Best in class manageability • Improved productivity through seamless management of PDW appliances by IT
Half Rack appliances	• Reduced acquisition cost for small to mid-sized MPP Data Warehouses
Multiple Physical Instances of Tables	• Fewer performance bottlenecks • Enterprise-class redundancy by mirroring all server and storage components
Massively Parallel Processing with Shared Nothing Architecture	• Better scalability, better and more predictable performance, reduced risk, and a lower cost per terabyte • Ability to partition data across nodes to reduce bottlenecks and contention
IO and CPU Affinity within SMP Nodes	• Eliminated contention per user query • Increased process performance using full resources of the node for each query

V. MASTER DATA MANAGEMENT

The management of the data involves provision of data storage and access to the user. In normal data warehousing environment such as a large data warehouse repository or individual data marts are there and data processing is done by those single repositories. Here important thing is availability of the same data in data warehouse repository and data marts. The solution is simple if the users are common to data warehouse repository and data marts then we need to replicate the data in both, otherwise no need to store the same data in both places. Name given to common data usage by data warehousing repository and data mart is known as Master data[5]. The concept of master data basically observed in the context of dependent data marts. As the data mart construction is done by data warehouse repository that will be the master data for data marts. In case of independent data marts the context is different. Depending on the request made by the user the vital information of the organization may be used by independent data marts on that case the master data is one which is recorded earlier by the organization so as to serve all the basic needs of the company. In either dependent or independent data marts some critical information is there such as Project details, Employee details, Clients information. The observation of relationship between project and employees or project and client or employee and client is obvious for better business processing. Such critical information may be master data. In many cases the critical information is not properly maintained which leads to inaccurate, inefficient results may cause inconsistency of the data which causes negative to the organization in the market .So the obvious solution to avoid inconsistency is maintenance of critical information of the organization effectively with out compromising. The critical data maintenance in effective manner is possible through master data management.

Master data is the high-value, core information used to support critical business processes across an enterprise; it is at the heart of every business transaction, application, report, and decision. Exploring, defining, and agreeing on why and where business users rely on master data across business processes and IT systems is one way to approach Master data management. The use of definitional techniques will help us separate master data from application-specific data because it allows us to quickly locate master data elements. One such technique successfully employed by many organizations is to first identify an organization's core business entities, in terms of parties (customers, employees, vendors, suppliers), places (sites, locations, offices, regions), and things (assets, products, services).Some important considerations we need to observe in the master data management are as follows.

a) Whether we are processing critical business transactions

b) Is the data in our core business entity created and managed in multiple systems?

The first consideration gives the details about profitability, efficiency but the second one is complex because the multiple system data maintenance is tedious. With the usage of parallelism in the core business entity we can achieve better performance in the form of faster data processing and accurate data management. If the data in our core business entity created and managed in single system then sequential processing is enough, but the data is created and managed in multiple systems through parallelism we can expect better data processing. To achieve this we proposed one method of identifying core business entity in multiple systems is through dominating sets in the nodes of the data warehousing environment.

The most important and influential thing in the maintenance of the data in data warehousing is incremental data movements towards real-time operations. Here real-time operations refer to data movement from one point to another instantly whenever there is a need. The data movement should be done in seconds. Yet, the data itself is of low value or trust unless it has been cleansed, enhanced, and transformed to fit the target purpose. Instead of delivering bad data faster, user organizations should include Data quality functions that improve real-time data as it's in transit. Achieving Data quality for real time data is very much tedious; the following are various factors to achieve quality.

1. Standardization is a way of normalizing all the data recorded.

2. Validation makes sure that the data entered by user is correct because through telephone or web interface the data is recorded.

3. Detection is used along with validation so as to identify unauthorized faulty usage of the data.

4. Classification is used to estimate the track of the users and to maintain the system according to the usage of customer.

5. Augmentation. A common form of augmentation is to append demographic information to a customer record; this helps to complete a 360-degree view of each customer.

One more factor we proposed to improve data quality in real-time data is Synchronization of customer data and their requests [10]. The main benefit o embedding this factor into

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

real-time data is possibility of crating customer groups who are having similar queries and similar data processing criteria. The way of achieving this synchronization is the customer query processing is monitored while implementing the transaction and in future if any similar data processing is done by the transaction then the monitor will create one log record and it will help to identify the similar data processing requirements of many users belongs to various locations. The implementation of monitors while execution of transaction in a single server/client is simple because the monitoring of that individual system is enough, but the monitoring of data processing while the transactions are running in multiple servers/clients requires a more sophisticated methodology. The methodologies we provide here is creating log file as that of single system and compare the data processing requirement of the customers if any similarity is there then create similarity-log with in the system. The same procedure is continued for all the systems which are currently processing the user requirements. So each system is now having similarity-log content .Now the synchronization process is applicable to similar log record between the number of systems and we need to make a new similar-log record which will have the similarity measures of all the participated systems which as scattered geographically. Checking out individual log-record and synchronize with other participated systems is also important, because there may be a possibility of similarity of log-records which will need to create one more similar-log record. The same process will be continued for all similar-log records, individual-log records existing in multiple systems. The entire above mentioned process may be implemented in parallel mode so as to improve the performance of the system in case of time and resource utilization.

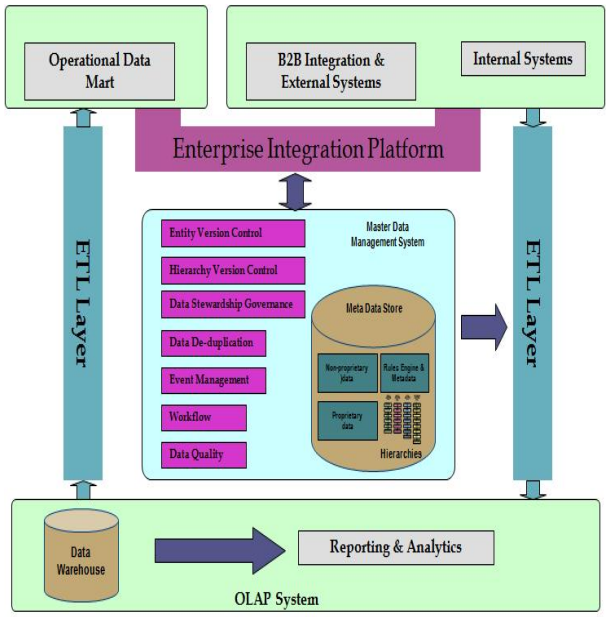


Fig. 5 Example of master data management

The importance of MDM can be observed from the following statistics. This study was done before the onset of the current economic downturn, and shrinking IT budgets are likely to take a toll on enterprise implementations, as IT budgets mirror the decline in business spend. All the same, global corporations need to stay competitive and prepare for the future. A significant cause of the sub prime crisis was insufficient assessment of risk and inadequate information sharing between lines of business in the enterprise. This occurs because information exists in multiple disparate systems and the processes are not interlinked, leading to inefficiencies and lost sales. Little surprise then that the Forrester report indicated that enterprises are looking to implement effective collaboration strategies, leveraging service-oriented architectures and master data management. These technology implementations are critical for business users to be able to get a single view of trusted data and ensure that there is a consistency in the way different units implement business processes. Companies that focus on achieving a high degree of data quality and standardized master data will be better able to deliver the right product to the right customer and at the same time ensure effective compliance and regulatory reporting. The end state of a successful master data management (MDM) initiative is a cleansed, real-time repository of master data like customers and products. This repository is updated from multiple source systems and publishes its information to downstream systems like data warehouses and operational systems, leading to closed loop data quality.

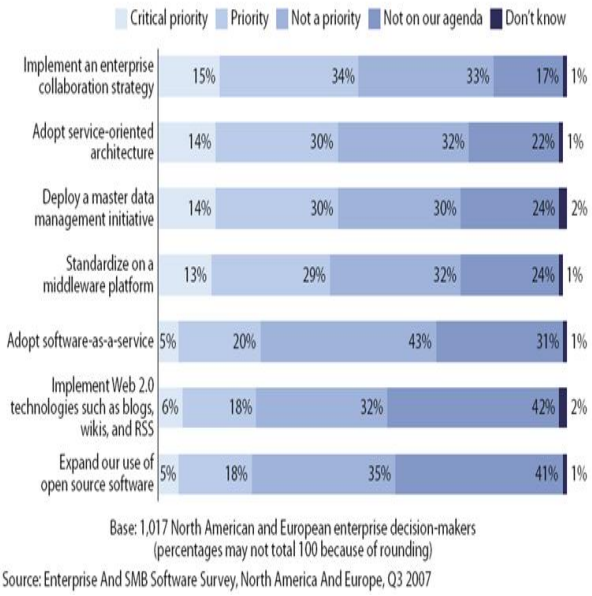


Fig. 6 Statistics of mdw usage

VI. DATA VIRTUALIZATION

Virtualization allows transformation of a server for multiple applications. The usage of resources we can observe in the following manner in case of stand-alone application and in virtualization. The above diagram depicts that the usage levels by dedicated application and in virtualized applications, it is clearly observed that the maximum possible resource utilization is done in virtualization [6] . So we can use this

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

virtualization concept in DWH environment and as well as PDW environments. Virtualization is an industry-changing movement that will touch all aspects of IT infrastructure and drive new levels of flexibility and dynamism in IT. Virtualization is addressing the process and operational issues around deploying and managing a large-scale virtual environment. Virtualization is based on Insertion of a hyper visor on Top of Hardware. Hyper visor installs immediately – Supports Desktops and Laptops. Virtual Machines Run on Any Hardware Configuration. Virtual Machines Can Run on a Shared Infrastructure. Single Software Can Span Different Hardware Components. Virtualization Allows Moving Applications without Service Interruption. The hyper visor is placed in the following manner in the systems.

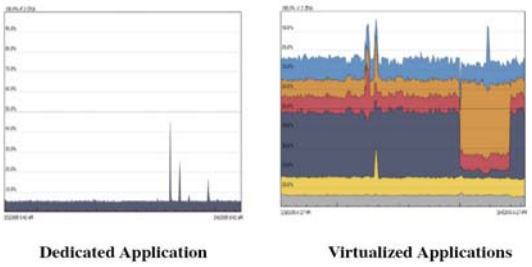


Fig. 7 Comparison of dedicated application with virtualization application

Some of the advantages of virtualization for effective maintenance of the data is as follows.

- Zero downtime maintenance
- Pooling hardware resource
- Virtual hardware supports legacy operating systems efficiently
- Dynamic resource sharing
- Freedom from vendor-imposed upgrade cycles

The following statistics identifies how virtualization will improve the performance of the system.

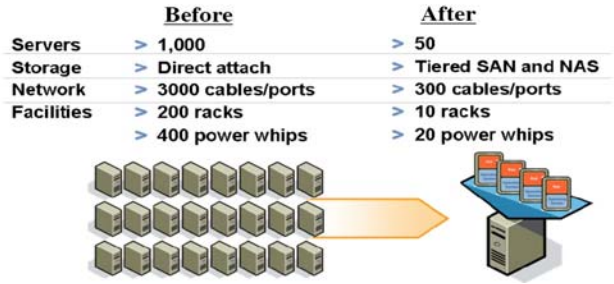


Fig. 8 Effective resource utilization in virtualization

The traditional view follows the methodology which is having Customer Relation ship management (CRM), VirtualPrivatenetwork (VPN), Exchange of data between various systems or users and File/Print option of the processed data [7]. But the following diagram depicts that usage of virtualization

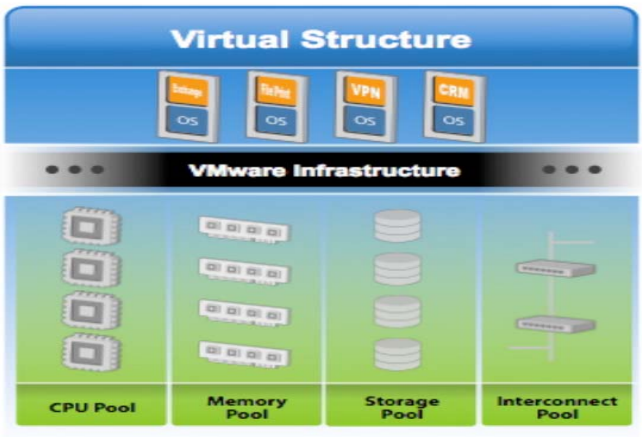


Fig. 9 Representation of virtualization in current systemisation

Through Virtual structure we can observe pool of data in the form of CPU Pool, Memory Pool, Storage Pool, Inter connect Pool, which will greatly helps the usage of cpu cycles, effective memory utilization, and provides the better storage environment to the user data and the provision of interfacing activity between the systems or users is done. Virtualization offers major savings in data centre Operations. Virtualization makes possible significant reductions in the costs of managing data centres, with simplification of systems management tasks.

Virtualization offers back-up and increased redundancy for delivery of high performance and high availability services. Virtualization is a step in the direction of “cloud computing”.

VII. MIDDLE-WARE TECHNOLOGY

Middleware Architecture defines the functions that enable communications in a distributed system and the tools that improve the overall usability of architecture made up of products from many different vendors on multiple platforms. Middleware is software that allows organizations to share data between disparate systems that do not communicate easily. Middleware has been described as the software “glue” that ties different applications together. Open source software has proven its value in companies of all sizes. As large enterprises have become increasingly comfortable with the open source model.

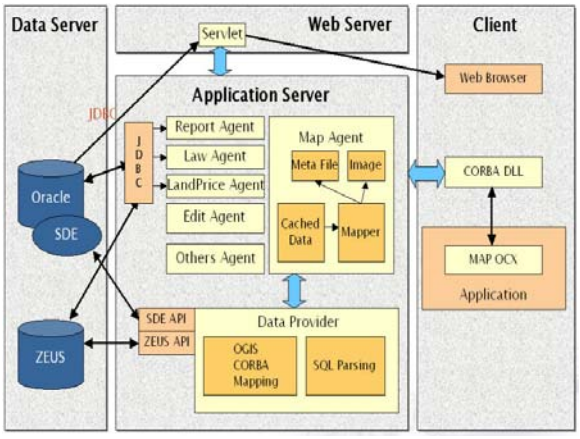


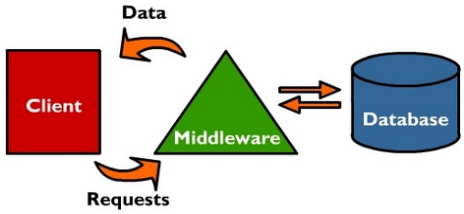
Fig. 10 Middle-ware technology usage in systems

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

VIII. MIDDLEWARE USAGE IN DWH

The companies with data base and data warehouse projects are now trying to adopt MWT to DWH and PDW.They have also begun adopting open source middleware for developing and deploying mission-critical applications. These encompass not only operating systems, but also middleware for application hosting, application integration, and data integration, including the development tools, management tools, and monitoring tools to fully support the stack [8]. When multiple databases must be used together, organizations need to bridge a number of different technical and semantic gaps to get the data they need. Organizations have increasing data volumes, increasing needs for consolidated data views to drive real-time business operations, and increasing drive toward interoperability and standards support. The ability to bridge data gaps in a more straightforward, streamlined, and scalable way is becoming an urgent need.

As organizations continue the drive to become more service-oriented, there's an associated drive toward improving the consumability of applications and services for end users. Lines of business will not be satisfied with the cost of SOA investment without enhanced end user services to achieve business objectives. They expect that reuse and modularity will occur at all levels, from site-wide to the component level, enabling the delivery of more personalized business applications. The population of power users within organizations will significantly expand as more technically savvy workers enter the workforce. This requires a new generation of user interfaces that allow power users to rapidly configure custom, time-sensitive dashboard applications without the need for an IDE or programming skills. This open source reference architecture is a maximally flexible, multi-purpose architecture. As with other reference architectures, it can provide a template for evaluating technology choices in each of the focus areas that have been highlighted, including application and service runtime, process management and service integration, data integration, and user interaction services [9]. The MWT in case of DWH environment is observed, but in case of PDW the things are changing as multiple users and multiple servers and multiple threads are activating simultaneously. The proposed work concentrates on how the data is accessed and processed by the users and systems to easily maintain the communication mechanisms with out contention of the network and smooth handling of the data processing we gave one framework .The framework first identifies the systems are selected by the users and other systems for data processing, we can make these systems active and try to find out the shortest paths from the users/system to system/users. By doing so it is possible to estimate the workload of the systems if any of the system is beyond the capacity of its ability then the request is redirected to the other system which is having fewer loads. The same thread will be running in all the communications so as to maintain the communication with out any problem to systems and users. To achieve the above mentioned framework there is a requirement of dominating sets which will identify the major participating systems in the data processing, as a next step we can redirect the data processing functionality to other system by observing the loads of the nodes(systems) in the framework. The final outcomes of this framework is identification of isolated systems and avoid those systems interaction in the current data processing, identify the systems with relevancy to the data processing but having fewer loads and redirect the data processing logs to these systems from overloaded systems. Middleware should provide flexibility, portability, and cost effectiveness in the implementation of enterprise architecture. The usage of MWT in PDW requires load balancing, synchronization, message queuing, failure recovery, and efficient transaction processing monitor management.



IX. CONCLUSION AND FUTURE WORK

In the introduction of this paper we described the concepts of DWHenvironment PDWenvironmentMDM, Virtualization and MWT and we described how these aspects are interrelated and how interoperation between these aspects is needed for the users. In the MDM how the critical parts are identified in the DWH environment and in case of PDW it is very complex, data virtualization is another important constraint to allow the data access from the systems from various users located geographically, this concept also requiring the critical transaction identification, through which we can able to establish Virtual Machine ware to avoid problems in data accessing. The other aspect we described is MWT in DWH and PDW, the usage of MWT in DWH is in the normal format such as Remote procedure call and CORBA but we suggested the open source MWT to handle PDW because the users are at various levels and they are having different levels of knowledge in data processing so usage of open sources such as C, C++ are better in those cases. The common conclusion is estimation of load between the users and systems and makes use of dominating set property and through that the entire processing of the DWH and PDW in the context of MDM, Virtualization and MWT. The future work for this research is implementation of critical sections in dominating sets for the DWH and PDW environments for Cloud technologies and high speed networks.

ACKNOWLEDGMENT

The ideas presented in this paper have developed over the years as part of the data warehousing development at various research communities. They have been refined through discussion with a number of people who have been affiliated with various research projects over the years. Specifically, we would like to acknowledge the contributions made by Dr.S.Saraswathi Associate Professor, IT Dept, Pondicherry Engineering College, and Pondicherry.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] B S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., G.N. Purohit and Usha Sharma, International journal on applications of graph theory in wireless ad hoc networks and sensor networks,(GRAPH-HOC) Vol.2, No.3, September 2010

[2] o Gerasimos Marketos, Mobility Data warehousing and mining, publisher, ACM. VLDB '09 PhD Workshop, August 24, 2009, Lyon, France.

[3] Anindya Datta,A case for Parallelism in data warehousing and OLAP

[4] José Blakeley White paper on Database Systems Group, Microsoft Corporation SQL Server Parallel Data Warehouse: Architecture Overview January 2010

[5] Lynne A. Dunbrack ,Program Director, Connected Health IT Strategies, Multidomain Master Data Management: Accelerating Speed to Value December 2011

[6] David S.LinthicumLLC,Next-Generation Data Virtualization Fast and Direct Data Access, More Reuse, and Better Agility and Data Governance for BI, MDM, and SOA

[7] Introduction to virtualization, Paul A. StrassmannGeorge Mason University, October 29, 2008

[8] Open Source Middleware Reference Architecture Leveraging JBoss Enterprise Middleware across your Enterprise to increase productivity and reduce costs

[9] The COTS Enterprise Architecture Workgroup, Middleware Domain Team, Middleware Architecture Report, A Middleware Framework for Delivering Business Solutions, Version 1.0, May 2001.

[10] Intel and Oracle on Dell PowerEdge Servers: Building Next-Generation IT, July 2010

[11] SQL Server 2008 R2 Parallel Data Warehouse, 2011 Microsoft Corporation.

AUTHORS PROFILE

Kethavarapu Uma Pavan Kumar pursuing PhD from Pondicherry Engineering College in Computer Science and Engineering from Pondicherry Engineering College under the guidance **Dr.S.Saraswathi, Associate Professor, IT Dept, PEC** received his M.Tech in computer science and Engineering from JNTU Kakinada University. His research areas of interest are Databases, Parallel processing, Data warehousing and mining.

Dr. S.Saraswathi is Associate professor, in the Department of Information Technology, Pondicherry Engineering College, Pondicherry, India. She completed her PhD, in the area of speech recognition for Tamil language. Her areas of interest include speech processing, artificial intelligence and Intelligent systems.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Talking Business Card Using Augmented Reality

Farimah Ghazaei
Master of Computer Science, Multi Media
Faculty of Computer Science and Information Technology
University Putra Malaysia, Serdang, Malaysia
Farimah.ghazaei@gmail.com

Sahar Sabbaghi Mahmouei
Master of Smart Technology and Robotic Program
Institute of Advanced Technology (ITMA)
University Putra Malaysia, Serdang, Malaysia
Sabbaghi.sahar@gmail.com

Abstract*—Augmented reality (AR) is a relatively new technology that allows mixing virtual with real world in different proportions to achieve a level of immersion that no virtual equipment can provide. Recent advances in the field of computers and virtual environments make possible AR technology to go many applications. AR technology aims to enhance the user's perception and interaction with the real world by implementing the real world with 3Dvirtual objects, which appear to coexist in the same space as the real world. The traditional business cards are no longer popular, since showing all the relevant data in a small space on a business card is impractical, time consuming or costly. Therefore there is a need for developing, an Augmented Reality Advertising Application which is capable of storing and representing a huge amount of data in a reasonable time and cost. This paper, implements the current status of the AR systems for Business Card which is the new type of automated applications and act to enhance the effectiveness and attractiveness of marketing for people in a real life scene. Augmented Reality Business Card (ARBC) is a really great way of getting people talking in something really special. This research have proposed a system which is based on Augmented Reality and designing a business card with marker.

Keywords-component; Business card; Augmented Reality; Marker; virtual environment.

I. INTRODUCTION

Showing all the persons information in business card has always been a problem. Mostly, imagination exactly the person's profession and details about them from the business card is not simple. Imagination 3D model from 2D Card depends on the capacity of the costumer to extract correct information from the Business Card. Besides, in order to explain about people' job in 3D images, it is needed to slice the images. Explaining about the Business Card with the help of text and images is quite difficult in complex situations whereas 3D sight of compound images raise the problem of understanding. [3]

Beside the problem of understanding and imagination of 3D objects, issue of motivation is very important, these days people are not satisfied with the quality of business card as they are not interesting enough. By offering a method which can interact with the user to observe live direct or in direct of the business card which is in real-time and in semantic context with environmental elements user can have a situation to experiment a virtual video in real world. Augmented reality (AR) is one of the newest technologies explored in business, promises the potential to revolutionize interface and the way of advertising, making consumer's experience more engaging. Recently, with the help of Augmented Reality it becomes more possible to improve the business environment to increase the quality of Business Card. AR allows you to create a more impressive business card at no cost at all and you can fit way more info on it. Business card design will never be the same with Augmented Reality (AR) letting you put more information than ever before on your pocket-size resume.

Augmented Reality will be one of the platforms in order to move towards a developed business and have an important role in marketing. Power of operating in different filed of business and manufacturing is one of the most significant advantages of this system.[3],[4].
AR has characteristics that make them suitable for business and can fulfill the lack of introducing the profession and business for businessman, such as:

- Remove the problem of limited space
- Disable the poor layout skills culminating in a usual card
- Allowing consumer to feel a sense of presence in the virtual environment
- Remove the problem with poor paper quality

Beside the characteristics of AR environment that suits the business environments, the growing research into the area of augmented reality in business is evidence that proves the great potential of this technology. The idea of create business card based on AR system started by James Alliban. The AR system including a physical card which is used to interface reality and virtuality and presents the human video in three dimensions (James Alliban). His work was based on AR technologies to enhance the business quality to increase the motivation of marketing by introducing through entertainment.

This study is focusing on proposing the AR technology on Business Card, which is based on Video see-through AR; rely

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

on video feed for acquiring information, uses webcam. With using marker it can register for virtual environment. In the next step it can display 3D video on top of the marker. This model help businessman to identify, adapt and implement Business Card to achieve their financial and marketing targets consistent with their strategic goals. In this study users are not allowed to add their own information or video.

II. PROBLEM STATMENT

Traditional business cards have helped businesses grow, establish relationships with affiliates and partners and also spread the word about their businesses and products. But, many of these business cards are boring to be honest. Most if not all, just show the giver's name, company affiliation (with logo) and how you can be contacted with them. Because typical business cards come with aspect ratios of dimensions range therefore it is unable to allocate more information due to the limitation of space. And also, sometimes the information that included in a business card is unclear or incomplete for its viewer. In aspect of economy, these traditional business cards are very costly. So, what if there was a very useful technology that we can use that for overcome to these problems. For example a professional business card, which could include one or more aspects of striking visual design such as map of the address location. Hence, there is a need to find a way to solve such problem.[1]

Today, with help of Augmented reality technology why waste your money to printing out your business cards when you can 'hand it out' for free? With the task at hand, this project is to design and develop a business card application, by using Video see-through AR; rely on video feed for acquiring information from webcam and interaction through the camera and markers. The idea of this application is to display clearer information by AR technology within the limitation space in physical object of real world for enhancing the functions of traditional business cards. [2]

In step of identifying the marker, the identification is not fast enough due to need of image matching with a library. In some projects marker detection techniques is based on matching the pattern on marker with a database or library of patterns, which is computationally expensive and is not fast enough. Whereas ID based markers can enhance the detection speed. In some AR application we see that, could not use the texture on their objects, and use the color instead although visualizing some elements without the texture is not efficient. But using Open CV library we can render the texture as well. [4]

III. OBJECTIVES

The main goal of this study is to develop a system for business people based on AR techniques in order to introducing their business which showing all required data while consumes less cost and time. The following are the objective of the proposed study by developing the Augmented Reality model:

- To design a business card with marker.
- To create a video of the business people.
- To detect and identify the marker to find 3D position, orientation and video to be viewed.
- To occlude the video onto the marker.

This study also could be very useful for costumer in their decision-making, company in their identifying process and identifying potential opportunities. It will be helpful for consumer to find the correct marketing among the entire product in interesting way. They can enjoy the business trade by technological innovations.

IV. ADVANTAGES AND DISADVANTAGES OF AGUMENTED REALITY

A. Advantages

1) The New Sphere:

As a result of creating Augmented Reality a new sphere has formed known as 'The Virtual Sphere'. This has produced a new platform for media to work with including in the Public Relations field. New campaigns are beginning to include Augmented Reality as part of their communications strategies. A recent campaign which has incorporated The Gorillaz for their new album, "Plastic Beach". They have promoted it in the latest edition of NME Magazine which comes complete with an A5 booklet filled with Gorillaz information and inside is an Augmented Reality marker, which when held up to a webcam the user is presented with a 3D "Plastic Beach" which may be navigated around. See Fig 1.[3]

Figure1. NME magazine which promoted by Agumented Reality marker

It could be argued that interactivity of this nature works well in PR campaigns and Augmented Reality is the next generation of interactivity with your consumers.

2) Personalization:

Personalization is the other aspect of Augmented Reality advantages. The concept of uploading people's media, such as the videos, helps to create a highly personalized piece of media for the user. It relates specifically to them, which is likely to be far more engaging than a standard video or image.

B. Disadvantages

1) Privacy:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Nevertheless, all of this causes great concern for the privacy of its users and realistically Augmented Reality cannot come without its drawbacks. The main problem with privacy for PROs is that there are no clear boundaries when it comes to accessing consumer's personal information.

2) Appearance:

Appearance concerns what the virtual objects look like.

V. Introduction of Business Card

Business cards are small cards, which are printed with business information about a company or individual. Business cards help one in giving great impression to stranger where it creates a professional and memorable impact every time. A typical business card includes the giver's name, company affiliation and contact information such as addresses, contact numbers, e-mail addresses and etc (Figure 2). [1]

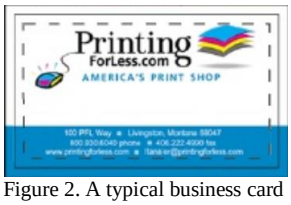


Figure 2. A typical business card

Nowadays, a professional business card includes one or more aspects of striking visual design such as map of the address location. The aspect ratios range of business cards is from 1.43 to 1.8(PrintingForLess.com, 2009). This is because different countries have varied standard for the size of business cards.[3]

A. Agumented Reality in Business card

Many business cards are boring to be honest. Most, if not all just show name, rank and how can be contacted. With using augmented reality business cards can be created. The cards were printed and on the back contain a graphic, which is then captured on to a computer via a webcam. When the camera picks up the little graphic a cool 3D grid of colored planes pops up, each extruding towards the camera depending on the brightness of the pixel, and then a video can be played back with the business card owner being seen talking with a message about businessman.

B. Current researches on augmented Business Card

Augmented Business Card is a business card that comes with marker for AR technology tracking. In simple words, this business card is exactly a typical business card with the implementation of AR technology in it. Some researches have been done in implementing AR into a typical business card. One example of the research is Toxin Labs' research where AR technology used in business card to display status of social networks (e.g.Twitter), show personal portfolio, calling or direct contact through the application and etc.(Jäger, 2009).[9] See Fig 3.

Figure 3. Agumented Business Card (by Jonas Jagar)

James Alliban (2009) also done such research, which developed as an interactive Augmented Reality application for businessman. He records a short video and created a 3D grid of colored planes. (Figure 4) [8]

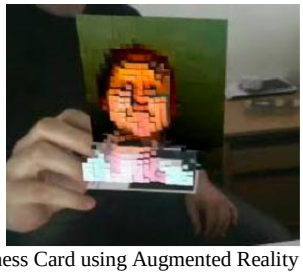


Figure 4. Business Card using Augmented Reality (James Alliban)

Burton Posey (2009) creates a business card which is worked in augmented reality. [10] In this card Burton Posey uses the avatar in the video that must be occlude on top of the card. He believes that the augmented reality avatar concept for the business card provides a way to provide potential clients and associates a way to get past the logo and brand and get to know the people that make the brand happen.

Figure 5. A virtual avatar on a business card by using Augmented Reality. (Burton Posey)

According to Wagner and Schmalstieg (2009), the marker is usually expected to be surrounded by black borders and containing a black and white pattern or image. However, the marker's border can be made very thin (down to zero) if the image inside the marker is darker (Wagner & Schmalstieg, 2009). This is because greyscale is converts from tracker internally.

C. Comparison each case

In this part all cases compare together and finally compare with this study to find out what they used in their projects.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1 shows the comparison between case studies.

Table I. comparison between case studies.

Case	Augmented Reality	Physical card	Animation	Sound	3D	Video
James Alliban	✔	✔	✔	✔		
Jonas Jagar	✔	✔	✔			✔
Burton Posey	✔	✔	✔		✔	
Current study	✔	✔		✔		✔

VI. HARDWARE AND SOFTWARE REQUIRMENTS

The minimum hardware and software requirements for Augmented Reality card are determined as below.

A. *Hardware Requirements:*

Hardware Requirements	
Component	Properties
processor	Intel Pentium 4 3.0 GHz or higher
RAM	1 GB of RAM or higher
Hard disk space	80 Gigabytes and above
Monitor	High resolution monitor
webcam	2 megapixel or higher

B. *Software Requirments:*

Software Requirements	
Component	• Properties
Operating System	• Windows 7
Adobe	• Adobe Flash CS4 • Adobe Photoshop CS4 • FLAR toolkit

VII. METHODOLOGY AND IMPLEMENTATION

In this section, the architectural design of the application developed is discussed. The language for the code is the action script in Adobe Flash CS4. Action Script enables us to efficient programming of Flash Platform applications for everything from simple animations to complex, interactive interfaces.

BC is an Augmented Reality Card, which can be used by cheap equipment including a normal PC and a webcam. In Flash this is usually done with a webcam and a marker. FLAR Toolkits the name of the Library, which is used in order to implement AR in this project. In general, FLAR Toolkit is software library for building Augmented Reality applications in Flash. The initial step for developing this project is taking video and changes the suffix to .flv and then use in coding part. After that regardless of which platform we are using we should consider configuring paths to the FLAR Toolkit library files. FLAR Toolkit library first version was released in May 2008 which is the world's first Flash based augmented reality library ported from NyAR Toolkit (Java ported version of AR Toolkit), to setup the FLAR Toolkit we should download the FLAR Toolkit library and the header files. To get the FLAR Toolkit library and the header files we need to download the FLAR Toolkit from [5],[6],[7].

Implementation is the process of moving an idea from concept to reality. In business, implementation refers to the building process rather than the design process. A result of implementing of this study is a finished system. Therefore all the processes to achieve the objectives including: BCAR design, creating video, creating marker, coding in action scripts, testing and debug, delivery.

A. *Process Modeling*

This part proposed the overview of this study. The video stream from marker is captured then video are rendered in video frame. With using T1 transform video to align them with markers, next the symbol inside of the marker is matched with templates in memory and positions and orientation of markers relatively to the camera are calculated. The image is converted to binary image and black marker frame is identified. Finally, video are rendered in video frame.

B. *BCAR Design*

To using the card as a user interface business card have been prepared. The card contains information about the business people profession, marker, and background for business card which is design using Adobe Photoshop.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 6a. Semantic view of in front of the card

Figure 6b. Back side of card

C. Creating Video

Video of business people captured using camera while they can explain about all aspects of their activity and their profession due to their need on advertising. Then it must be converted to .flv to be suitable in this project. For creating the video simple digital camera with high resolution is needed. The short video has taken in this project, which is around 10 minutes.

D. Creating Marker

The system was developed using fiducial markers, the fiducial markers are captured by webcam while the system verifies and identify the marker. Marker based AR uses a camera and a visual marker known as a fiducially to determine the centre, orientation, and range of its spherical coordinate system. Marker is designed for detecting the video form camera and displays it on the screen. The Flash Augmented Reality Toolkit library holds much of the magic for this application. It will introspect the webcam image for a particular graphic image and judge where in real space to map the 3Dmodel. See Fig 7.

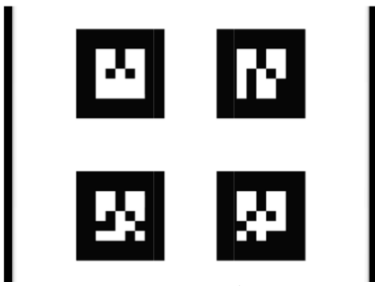


Figure 7. Sample of marker

The FLARCameraParams.dat file, which comes with the toolkit, will be used as-is. It corrects webcam distortion and helps the toolkit with its general detection routine. The graphic marker image information is held within the FLARPattern.pat file. The marker in this project is designed by Adobe Photoshop in black and white color with 800x800 pixel file with 150 pixel border around the picture. With the help of online marker creator the marker convert to .pat file. Then it can be recognize and capture by camera.

Bitmap image (.pat file) used as a pattern that can be imported into multiple graphics programs. Typically a square file that may be 8x8 pixels to 256x256 pixels in size which is often used for creating a textured background. In online marker detector set the marker segments to 16×16 and the marker size to 50%, show marker to webcam and when a red square surrounds marker click on get pattern, then click save.

Figure 8. Online Marker Detectors

To detect the marker and load it from the folder has used some codes.

```
public function MultiFLARExample() {
    this.stage.align = StageAlign.TOP_LEFT;
    _cubes = new Array();

    _markers = new Array();
    for(var i:int=0;i<30;i++)
    {
        video_status[i]=0;
    }
    //_markers.push (new FLARMarkerObj( "rabih.pat",16,50,80));//1
    //_markers.push( new FLARMarkerObj( "lili.pat" , 16 , 50 , 80 ) );//2
    //_markers.push( new FLARMarkerObj( "azreen.pat" , 16 , 50 , 80 ) );//3
    _markers.push( new FLARMarkerObj( "queen.pat" , 16 , 50 , 80 ) );//4
    super( );
}
```

Here is the code for the position of the marker, which is used in this study, defines to the video.

```
override protected function _detectMarkers() : void {
    _resultsArray = _flarDetector.updateMarkerPosition( _flarRaster , 80 , .5 );
    for ( var i : int = 0 ; i < _resultsArray.length ; i ++ ) {
        var subResults : Array = _resultsArray[ i ];
        for ( var j : * in subResults ) {
            _flarDetector.getTransmationMatrix( subResults[ j ], _resultMat );
            if ( _cubes[ i ][ j ] != null ) transformMatrix( _cubes[ i ][ j ] , _resultMat );
        }
    }
}
```


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

VIII. TESTING AND DEBUG

In this stage the program involves locating and removing any errors in the program. Testing and debugging step by step is the process of finding and correcting errors. The thing that is so important is that the expected results of tests should be appearing. The main objective in testing is to find errors. After we find the errors the changes should be apply in order to enhance system efficiency.

Testing can also be stated as the process of validating and verifying of system program. In this stage, the developed augmented reality system was tested among both academic and business sectors. Two lecturers from faculty of computer science, University Putra Malaysia and two business people have chosen for this test.

Figure 8a. Sample of Business card in academic sector

Figure 8b. Sample of Business card in academic sector

Figure 8c. Sample of Business card in business sector

From the results of the testing, it can be concluded that the developed business card was not interesting among the some of the academic staff which prefer to use the static business card, while the business sector considered it as an interesting approach in the business strategy. We believe that the reason that business card was not interesting among the academic sector it because the field of AR still new in this kind of research.

IX. DELIVERY

After the system tested and the favorite result achieved then it should be deliver to the user. In this project the EXE file and the sample of the marker putted in the web site, which all the user can easily run the project and it automatically will display on to the screen.

X. CONCLUSION

This project has argued that Augmented Reality is the best instrument to design the Business Card .The aim of Virtual Reality is to replacing the perception of the world with an artificial one; besides Augmented Reality has the goal of enhancing a person's perception of the surrounding world. Being partly virtual and real, the new interface technology of AR which is able to display relevant information at the appropriate time and location, offers many potential applications; these include aiding in business, selling and buying the belongings, repair or maintenance, manufacturing, medicine, game. The purpose of the current study was to determine the new method for introducing the business people profession.

The most obvious finding to emerge from this study is that Talking Augmented Reality Business Card support video on objects and motivates the consumer by allowing them to feel senesce of presence in the virtual world, where big company, employee, private company can used it. It can also use for modeling some virtual three-dimensional business card and the video will be created on top of a marker which is printed on a paper. Video will appear when marker is viewed through a web camera. AR is a really great way of getting people talking in something really special. The simple card is just can describe brief description about the people profession but the new method is introduced to overcome this issue. The Augmented Reality Business Card can be seen in an ample quality of advertising can be considered as a successful collaboration in purchasing activates.

Finally, the project begins development cycle to create interactive business cards so that people could customize their avatars, videos and exchange them. Since the subject of the project is a business conceptual, the real outcomes are supposed to be exposed in a business environment. The

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

evidence from this study suggests that it should be better to use the Business Card only in business environment. The methods used for this Business Card may be applied to other applications elsewhere in the world.

XI. FUTURE WORKS

The future research for this study can be seen in many ways. The most relevant subject could be development of this business card for the big company to reveal all the information about the situation of their employee's place in the monitor, which is design in the beginning part of the company of their tasks by displaying the right information at the right time and place.

There are many technical challenges to be overcome before such interfaces are widely deployed, but driven by compelling potential applications in surgery, the military, manufacturing, and entertainment, progress continues to be made in this promising form of human-computer interaction.

REFRENCS

[1] Billinghurst, M., H. Kato, and I. Poupyrev, The MagicBook: a transitional AR interface. Computers & Graphics, 2001. 25(5): p. 745-753.

[2] Chen, C. and J. Zhang. Design and realization of Computer Aided Instruction platform based on Augmented Reality:2008, IEEE

[3] Edmon NG,Te Tezong, B.Parhizkar and Arash H.Lashkari. Mobile phone agumented reality business card:2011. International journal of computer scinece and information security.

[4] Fiala, M., ARTag, a fiducial marker system using digital techniques. 2005.

[5] Kato, H., M. Billinghurst, et al. (1999). "ARToolKit." Hiroshima City University.

[6] Mark, F., Comparing ARTag and ARToolkit Plus Fiducial Marker Systems, in IEEE International Workshop on Haptic Audio Visual Environments and their Applications. 2005: National Research Council of Canada, NRC 1200 Montreal RD, Ottawa, Canada K1A-0R6.

[7] Wente, M.N., et al., Perspectives of evidence-based surgery. Digestive surgery, 2000. 20(4): p.263-269.

[8] http://jamesalliban.wordpress.com/2009/06/03/ar-business-card/

[9] http://adland.tv/content/jonas-j-ger-opens-augmented-businesscards-website

[10] http://www.burtonposey.com/interactive/augmented-reality-business-card

AUTHORS PROFILE

Farimah Ghazaei graduated as a Master student from Faculty of Computer Science (FSKTM), Universiti Putra Malaysia, UPM in 2011.
Farimah has received her B.Sc in software computer engineering field in 2008 from Iran Azad University of Mashhad. Her research interest includes Image Processing, Computer Vision, Augmented Reality and Pattern Recognition.

Sahar Sabbaghi Mahmouei graduated as a Master student in Institute of Advanced Technology and Research (ITMA), Universiti Putra Malaysia, UPM.
Sahar has received her B.Sc in software computer engineering field in 2006 from Iran Azad University. Her research interest includes Image Processing, Machine Vision, Artificial Intelligence, Augmented Reality and E-commerce.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Survival Analysis In Cancer Gene Using Vector Space Model

Jitasha Mishra
Assistant Professor,
Computer Science and Engineering,
Gandhi Institute Of Technology And Management,
Bhubaneswar,India
Sibu_124@yahoo.co.in

Debashis Hati
Assistant Professor,
Computer science and Engineering,
Gandhi Engineering college,
Bhubaneswar,India
d_hati@yahoo.com

Amritesh Kumar
Assistant Professor,
Computer Science and Engineering,
Gandhi Institute Of Technology And Management,
Bhubaneswar,India
amritesh.kiit@gmail.com

Abstract—**The study is an effort to design a stable classification system and make a survival analysis to categorize microarray gene expression profiles. Currently, high-throughput microarray technology has been widely used to simultaneously probe the expression values of thousands genes in a biological sample. However, due to the nature of DNA hybridization, the expression profiles are highly noisy and demand specialized data mining methods for analysis. Our proposed approach focuses on developing an effective and stable sample classification system using gene expression data. The traditional cancer prognostic tools of tumor stage and morphology are inadequate benchmarks for the accurate determination of patient risk. The emergence of microarray technology has enabled the simultaneous measurement of thousands of gene expression levels, allowing researchers to apply sophisticated data mining and statistical techniques in the search for a superior prognostic methodology. This paper extends an existing procedure called Bayesian Model Averaging (BMA) to Cosine Bayesian Model Averaging for application to survival analysis. Cosine BMA is a method for predicting survival prognosis by isolating a small group of relevant predictor genes from a high-dimensional microarray dataset. In this paper, the Cosine BMA algorithm for survival analysis is applied to two real cancer datasets: diffuse large B-cell lymphoma and breast cancer. The selected genes are used to divide patients into high and low-risk categories. Results show that the Cosine BMA algorithm for survival analysis consistently selects a small number of relevant genes while providing a higher degree of predictive accuracy than other feature selection methods. The procedure shows promise as a powerful and cost-effective prognostic tool in future cancer research.**

Keywords- Microarray, Supervised learning, survival analysis, classification, DNA, FNAC, biopsy.

I. INTRODUCTION

In the field of bioinformatics we implement a number of algorithms to solve the problems like cancer. A cancer is a disease that is a big threat to human life. But many a starving to overcome the problem to make the mankind simpler and smoother. The researchers have gone through a multiple number of solutions but it requires something that would be the optimal one. Bioinformatics has led to a vast amount of research advances and has proven effective for diagnosing, classifying and discovering many aspects that lead to diseases like cancer. The focus from a macro level to a molecular level has led to a better understanding of the functions of genes. Various developments in the field of bioinformatics have led to efficient data mining and classification algorithms and techniques. The present study involves the application of machine learning methods for the classification of cancer samples using the gene expression data obtained from the micro array experiment.

Understanding gene expressions: A brief explanation of gene expression and micro arrays will help aid in the proper understanding of the current classification problem. Genetic material is the same in all cells of the body. The only thing that makes the organs in the body act differently is that some genes are dormant in certain cells. Some genes are expressed in a cell while others are not, creating the whole variation. These dormant genes in the cell are sometimes triggered in some circumstances which lead to several diseases and disorders like cancer .This leads to malfunctions in the proper working of the cells. Bioinformatics research shows that the expression levels of genes away from normal samples might be a reason for several abnormalities. With the help of new technologies, we are now able to study the expression levels of thousands of genes at once. In this way, we can try to compare the expression levels in normal and abnormal cells. The expression values in affected genes can help us compare them with regular expression values and thus tell us the reason for the abnormality. A gene is considered informative when its expression helps to classify samples to a disease condition or not. All of these informative genes help us develop classification systems which can distinguish normal cells from the abnormal ones. The goal of this study is to build a classification model which can efficiently classify the normal and tumor samples using gene expression data obtained from micro array study

Understanding micro arrays: A micro array is a tool used to sift through and analyze the information contained within a genome. A microarray consists of different nucleic acid probes that are chemically attached to a substrate, which can be a microchip, a glass slide or a microsphere-sized bead. The first DNA microarray chip was engineered at Stanford University,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

whereas Affymetrix Inc. was the first to create the patented DNA microarray wafer chip called the Gene Chip. The microarray data used for the current study was collected using Affymetix Gene Chips also knows as an oligonucleotide microarray. Messenger RNA is extracted from the cell and converted to cDNA. After the amplification and labeling of the sample it is hybridized on the chip. After the washing of unhybridized material, the chip is scanned with a laser scanner and the image analyzed by computer.

The more accurate method by which to assess relatively recent emergence of microarray technology has allowed for new advances in the determination of survival prognosis through the study of gene expression levels. Our proposed approach presents a survival analysis extension to the Cosine Bayesian Model Averaging (CBMA) algorithm , which is a method for predicting survival prognosis by isolating a small group of relevant predictor genes from a high-dimensional microarray dataset. The purpose of our proposed approach is to show that the Cosine BMA algorithm for survival analysis consistently selects a small and cost-effective number of predictor genes while outperforming other feature selection, classification and supervised learning procedures in the accuracy of patient risk assessment.

II Background Work

The science of biology has proven to be one of the most important areas of knowledge in human history. A tremendous amount of data is required to explore biological phenomena, and computer science techniques have become pivotal in managing this mountain of information. Bioinformatics is the merger of biology and computer science[9]. Bioinformatics spans the complete range of biological topics, but the study of gene expression is one of the most prominent and important subjects to pursue in the quest for medical advances. Gene expression is defined as the process by which a gene's DNA sequence is converted into a functional protein [3]. The DNA sequence is first encoded as messenger RNA (mRNA) through the process of transcription, and the mRNA is subsequently translated into a protein [2]. A higher expression score denotes greater amounts of gene activity. Gene expression profiles are created and studied to determine which genes are expressed in particular cell types, at what times these genes are expressed, and under what conditions expression occurs. Medical researchers often use these profiles to compare the expression levels of normal cells with those of cells in a special condition. This condition could involve cancer, other diseases, starvation, extreme temperature, or any other unique cellular state that scientists are interested in understanding. By studying the differences between the expression levels of these cells, researchers are able to determine the effects of the given condition on the process of gene expression in a particular cell type [8]. Gene expression profiles are most commonly extracted through the use of microarray technologies. A single silicon microarray chip can measure the expression levels of tens of thousands of genes at once, and this number changes rapidly as the technology grows more robust. To put this in prospective, the human genome is believed to contain around 20,000 to 5,000 genes, so one microarray chip could ostensibly produce expression scores for more genes than exist in the entire human genome [2]. Microarrays come in several different types, including short oligonucleotide arrays, cDNA or spotted arrays, long oligonucleotide arrays, and fiber-optic arrays. Short oligonucleotide arrays, manufactured by the company Affymetrix, are the most popular commercial variety on the market today [2]. A growing number of studies have used a variety of statistical methodologies to perform classification high-dimensional microarray data [1]. Developed a risk index based on 50 genes that classified lung cancer patients into either low- or high-risk groups with a considerable degree of accuracy[4]. The two groups differed significantly from one another in terms of survival rates (p-value=0.0006), and these results still held among patients with stage-1 tumors (p-value=0.028). It tested their risk index through leave-one-out cross validation on the original data set [4]. The index was then applied to an independent dataset with the same genetic expression information. Again, the high- and low-risk groups were found to differ significantly from one another, both overall (p-value=0.003) and among patients with stage-1 tumors (p-value=0.006). some approach applied a similar methodology based on 64 genes, with improved results. The study used multivariate Cox Proportional Hazards regression with bootstrap resampling and forward selection to identify a 64- gene signature, whereas other approach employed univariate Cox regression [5][4]. In addition, one approach gathered seven independent data sets for validation purposes [7]. Once they created a risk index by training on two of the seven datasets, they divided patients from the other five into high- and low-risk groups. In all cases, the two groups were significantly different from one another ($p < 0.001$). Published a survival classification study involving primary invasive breast carcinomas [4].They attempted to classify patients into two groups: those whose cancer recurred within five years of diagnosis and tumor resection (poor prognosis group), and those who remained disease-free beyond five years (good prognosis group)[9]. They used a three-step supervised classification method to develop a 70-gene predictive signature, which they applied to an independent test set of 19 patients. Their classifier correctly labeled 17 of the 19 samples, an accuracy of 89.5% (p-value=0.0018).

More recently, some approach used a combination of three different feature selection methods to identify a set of predictive genes: Prediction Analysis for Microarrays (PAM), Significance Analysis for Microarrays (SAM), and a correlation-based technique similar to that of) [1][7]. Of the genes selected in each method, 21 were present in all three models. This 21-gene signature was applied to two independent acute myeloid leukamia datasets and successfully separated patients into good and poor outcome groups ($p < 0.0005$).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

III PROPOSED ARCHITECTURE

The proposed architecture includes the total structure of our proposed approach. Initially by using supervised learning the system is trained on trained dataset. Then a feature extraction is done by isolating a small set of cancer affected genes and a microarray data is constructed by using the mathematical exponential function. Finally the classification and survival analysis is done. A microarray data is fed as input to the risk ratio calculation system. Then a survival of cancer genes done.

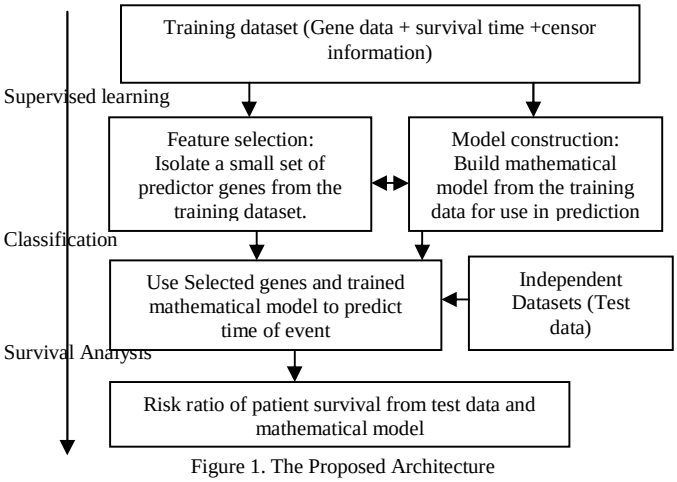


Figure 1. The Proposed Architecture

Steps involved in synthesizing a dataset

1. Collection of training dataset
2. Learning by a supervisor machine.
3. Transformation of training dataset to test dataset
4. Conversion and feature extraction.
5. Synthesis using mathematical model
6. Test conducted on test dataset.
7. Survival analysis on microarray data.

IV Proposed Approach

A. *Cosine Bayesian model averaging method for classification*

The Cosine BMA implementation described above is incompatible with microarray data. This is because the typical microarray dataset contains thousands or even tens of thousands of genes, but the leaps and bounds algorithm can only consider a maximum of 30 variables when selecting the top best models to return to the user. The usual practice of employing stepwise backward elimination to reduce the number of genes down to 30 is not applicable in a situation where the number of predictive variables is greater than the number of samples. An Cosine BMA algorithm that takes a rank-ordered list of genes and successively applies the traditional BMA algorithm until all genes up to a user-specified value p (G_1, G_2, …, G_p) have been processed. It begin by using the ratio of between-group to within-group sum of squares to rank-order the genes from the microarray dataset. As the algorithm iterates, genes with a high posterior probability are retained while genes with a low posterior probability are eliminated. The default threshold for inclusion is set to 0.3; genes whose posterior probabilities are less than 0.3 are discarded.

Breast Cancer: The breast cancer is a disease that occurs in human beings in a high rate. It is required to classify and cure it. The malignancy is present or not can be known by two ways clinically and pathologically. Various symptoms of breast cancer are presence of lump, discharge of blood, retraction of nipple and distraction of shape. Here we find the cancer genes of breast and classify them by the proposed algorithm and find the survival of the patients. The different values used for classification are:

Test	Total	Values	Result
MAMMOGRAPHY-BIRADS (Hormonal test for cancer)	100	Below -50	Normal
		Above 50 and upto 60	4
		Above 60 and upto 70	5
		Above 70 and upto 100	6
F.N.A.C. (Blood test for cancer)	100	Below -10,	Normal
		Above 10 and upto 100	Suspicious for Malignancy
BIOPSY (Tissue test for cancer)	100	Below -50,	Normal
		Above 50 and upto 60	BR Score-7
		Above 60 and upto 70	BR Score-8
		Above 70 and upto 100	BR Score-9

Table 1 Tests conducted for breast cancer

Diffuse large b-cell lymphoma: Dlbcl cancer is a type of cancer that occurs when the count of lymphoblast increases in blood. Lymphoblast is present in white blood corpuscles. It occurs many human beings so requires an approach for classification and estimation for curing. The different values for dlbcl cancer classification are:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Test	Total	Values	Result
MONOCLONAL BAND (Blood/Urine test for M-band, monoclonal band, monoclonal spike, monoclonal protein)	100	Above -0	CANCER POSITIVE
		Less than -0	NORMAL
ESR (Cell test in Blood for cancer)	100	Above & equal to 50	CANCER POSITIVE
		Less than 50	NORMAL
LDH (Tissue test in Blood for cancer)	1000	Above & equal to -500	CANCER POSITIVE
		Less than -500	NORMAL
F.N.A.C. (Blood test for cancer)	100	Below -10,	NORMAL
		Above 10 and upto 100	Suspicious for Malignancy

Table 2 Tests for DLBCL cancer

B. Cosine Bayesian model averaging method for survival analysis

In order to extend the Cosine BMA method to survival analysis, a number of algorithmic modifications were implemented. First, instead of applying the SS/WSS technique to rank-order the genes in the preprocessing step, the Cox Proportional Hazards is used to rank each individual gene. Cox regression is a popular choice in the realm of survival analysis due to its broad applicability and capacity for handling censored data. It is a semi-parametric method that quantifies the hazard rate for a subject *s* at time T as follows:

$$\lambda(T \mid p_s) = \lambda_0(T)\exp(p_s\theta) \qquad (1)$$

In this equation, λ_0 (T) is the baseline hazard function a time T, p_s is the vector of effect parameters for subject s and θ is the vector of unknown predictor coefficient. we observed that the baseline hazard function in equation (1) could be left unspecified if the effect of a covariate on one individual remains the same for all times T (e.g., if an environmental variable doubles your personal risk of dying at time 5, it also doubles your risk at time 8). Therefore, an estimation of θ is all that is needed. Our proposed approach calculate similarity between existing patient dataset and new patient dataset based on vector space model.

$$R(t,p) = \frac{\sum wkt * wkp}{\sqrt{(\sum (wkt)*(wkt))*(\sum (wkp)*(wkp))}} \qquad (2)$$

In equation (2), R(t,p) is the relationship between training dataset and test dataset, where training dataset is for existing patients and the test dataset is for new patient. The value of R(t,p) is in between 0 and 1. If the value of R(t,p) is above or equal to 0.3 then the patient is in high risk. wkt is the weight of the existing patient training dataset attributes. wkp is the weight of the new patient test dataset attributes. Where the patient attributes are diameter of the cell, texture, length, fluidity, cell size and volume for breast cancer and lym number, analysis set, follow up years, status, subgroup and germinal centre of b-cell in DLBCL cancer.

Following this step, the algorithm iterates through the user-specified p top-ranked genes, applying the Cosine BMA algorithm for survival analysis to each group of variables in the current BMA window (where the window size is denoted by x_{max}). This part of the procedure is similar to the classification method described previously; genes with high posterior probabilities are retained while genes with low posterior probabilities are eliminated.

To evaluate the performance of this method, the continuous risk scores of the patients were discretised into risk groups. The overall risk score for a single patient is the weighted average of the risk scores calculated for each model in the set S of contending models.

V. Proposed Algorithm (Cosine BMA)

The Cosine Bayesian model averaging method algorithm for survival analysis is used by me in my proposed approach for survival analysis of cancer patients is as follows:

STEP 1 Input: training set TD with G genes and n samples

STEP 2 Pre-processing step: Rank-order all G genes by applying Cox Proportional Hazards Regression to each individual gene. Let x1, x2, … , xG be the ordered list of genes, sorted in descending order of log likelihood. Let x_{max} denote the user-specified size of the Cosine BMA window (maximum 5).

STEP 3 Parameters: n best and p, where p is the total number of genes to be processed by the Cosine BMA algorithm and G is the total number of genes in the training dataset. Note that $x_{max} < p \leq G$.

STEP 4: Initially, start with the x_{max} top ranked genes (x1, x2, ., x_{max}), and apply the traditional BMA algorithm for survival analysis . Let to be processed be an ordered list of genes with ranks (x_{max} + 1) to p. Initially, to be processed
(x_{max} +1, x_{32}, …, x_p).

STEP 5: End

VI *Results and Implementation* :

A. Methods and Materials

Our proposed work is to find out the rate of survival of cancer patients. So, we have considered two cancer datasets for this proposed work. The two cancer genes involve breast cancer and diffuse large b-cell lymphoma cancer.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Attributes	Value
Diameter of the cell	0.38
Texture	-0.024
Length	-0.242
Fluidity	0.732
Cell size	-0.259
Volume	-0.134

Table 3 Training Dataset For Cancer

Attributes	Value
Diameter of the cell	0.092
Texture	-0.347
Length	-0.157
Fluidity	-0.04
Cell size	-0.575
Volume	-0.291

Table 4 Test Dataset of a cancer Patient

Here we are implementing table 3 and table 4 in equation (2), where the total value of training dataset is represented by wkt and the total value of test dataset is represented by wkp.

When we correlate all the values of table 3 and table 4 with equation (2), we find

0.239921/ 0.685752088= 0.349863463.

As this value is greater than 0.3 so this patient is suffering from cancer.

Diameter of the cell	Texture	Length	Fluidity	Cell size	Volume
-0.511	0.045	-0.737	0.226	-0.084	-0.15
-0.448	-0.573	-0.264	0.109	0.484	0.416
0.005	0.496	-0.377	0.435	-0.161	0.186
-0.097	0.242	-0.497	-0.639	-0.21	-0.715
0.046	-0.075	-0.832	-0.499	-0.518	-0.481
-0.15	-0.436	-0.786	-0.299	0.585	-0.663
-0.798	-0.506	-0.349	-0.591	-0.534	-0.671
0.854	-0.637	-0.775	-0.63	0.353	-0.319

Table 5 Dataset of breast cancer

In table 5 there are the test datasets of patients those who are suffering from breast cancer.

LYM number	Analysis set	Follow up years	Status	Sub group	Germinal center of b-cell
Length	Fluidity	Cell size	Volume		
1	0.5	-0.22	-1.29	-0.56	-0.3
3	1.15	-0.31	-0.14	-0.05	-0.68
5	1.72	0.59	0.79	-0.04	0.9
9	0.32	-0.05	-0.76	-0.66	-0.75
13	0.91	-0.04	-0.6	1.48	-1.32
14	0.43	0.23	-0.22	1.35	-1.19
16	1.44	-0.41	0.32	-0.7	-0.2
17	-0.16	0.71	-0.53	0.35	-0.21
19	-1.59	0.51	-0.38	-2.41	1.27

Table 6 Dataset of DLBCL cancer

In table 6 there are test datasets of patients those who are affected by DLBCL cancer.

B. Results

This section presents the results from the application of the Cosine BMA algorithm for survival analysis to the DLBCL and breast cancer datasets.

DLBCL cancer: In order to determine the best combination of these input parameters, a series of 10-fold cross validation runs were performed on the DLBCL training dataset becomes inefficient for BMA windows larger than 5 variables.

Classification for DLBCL cancer: In this classification the monoclonal blood sample test, ESR, LDH and FNAC tests are conducted by taking the blood samples of different patients. Then after carrying out the supervised learning the test dataset of a patient is taken to classify a patient is cancer affected or not.

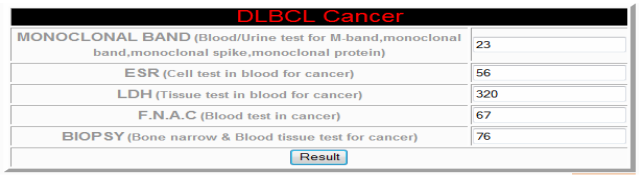


Figure 2

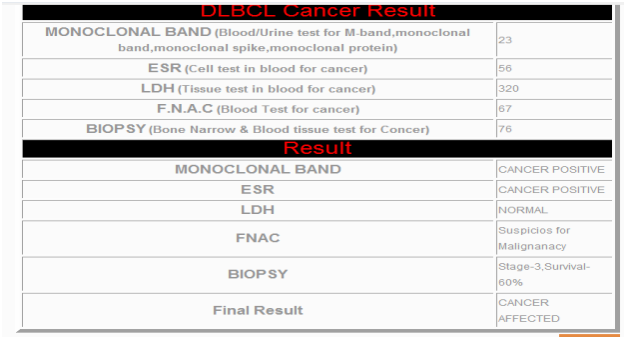


Figure 3

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

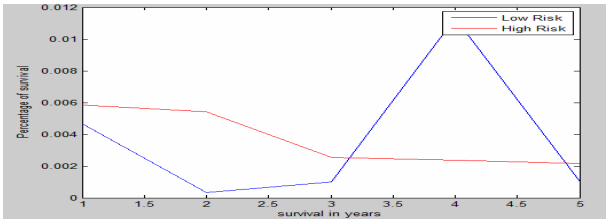


Figure 4: **Curve showing survival of DLBCL cancer patients**

Breast cancer: In the application of the Cosine BMA algorithm to the breast cancer dataset the optimal input parameters as determined from the cross validation study on the DLBCL dataset were used. During preliminary analyses training sets with fewer than 100 samples tended to result in fatal errors at higher values of maxNvar.

Classification of a particular patient: Firstly, I have considered the various values for breast cancer test are taken and learned to a system. Then a test data set of a patient is taken having original values of mammography birads, fnac test and biopsy tests that finally shows the survival chances on a scale of five year scale.

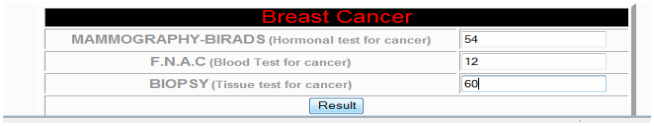


Figure 5

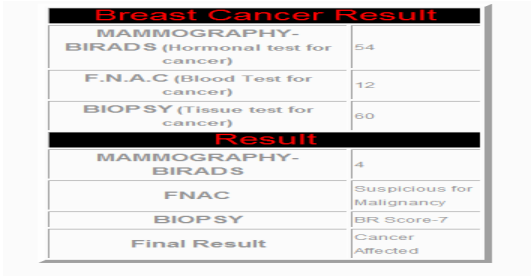


Figure 6

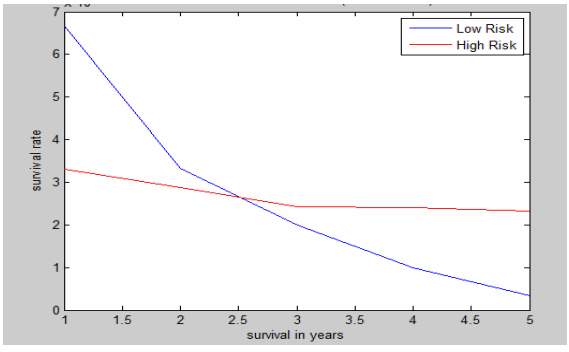


Figure 7 Curve for survival rate of breast cancer

VII Conclusion and Future Work

A. Conclusion

This paper has demonstrated that the Cosine BMA algorithm for survival analysis consistently isolates a small group of relevant predictor genes from microarray datasets, providing better predictive power than other feature selection and supervised learning methods. The algorithm is easy to use. It efficiently reduces vast amounts of microarray data down to a handful of predictor variables, making it a cost-effective diagnostic tool in the clinical setting.

B. Future Work

Future work in this area would involve collaboration with cancer biologists to validate the predictor genes selected by applying the Cosine BMA algorithm to microarray data, and to assess the prediction accuracy of the methodology on PCR data generated using independent patient samples. By implementing this technique on certain hospital we can cure a cancer patient in a better way. Furthermore, the Cosine BMA algorithm could be extended to other types of high-throughput data such as proteomics data produced from mass spectrometry. The multivariate nature of BMA combined with its ability to account for model uncertainty makes it a particularly attractive candidate to extract predictive genes from any high-dimensional biological dataset.

REFERENCES

[1] J. E. Korkola, E. Blaveri, S. DeVries, D. H. Moore, E. S. Hwang, (2007). "Identification of a Robust Gene Signature that Predicts Breast Cancer Outcome in Independent Data Sets". BMC Cancer, 7 (61),1471-2407/7/61.

[2] T. Golub, D. Lonim, P. Tamayo, C. Huard, M. Gaasenbeek(1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". Science, 286 (5439), 531-537.

[3] E.Bair, & R.Tibshirani (2004). Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. PLOS Biology, **2** (4), 511-522. .

[4] D. Beer, S. Kardia, C. Huang and T. Levin, (2002). Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. Nature Medicine, **8** (8), 816-824''

[5] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, (2000). Tissue Classification with Gene Expression Profiles. Journal of Computational Biology, **7** (3-4), 559-583

[6] L. Jiangeng, D. Yanhua, R. Xiaogang,(2007). "A Novel Hybrid Approach to Selecting Marker Genes for Cancer Classification Using Gene Expression Data". The International Conference on Bioinformatics and Biomedical Engineering, 2007, ICBBE 2007, 264-267

[7] C. Chen, Y. Huang and J. Lee, (2003). Characterization of the Univariate and Multivariate Techniques on the Analysis of Simulated and MRI Datasets with Visual Task.Nuclear Science Symposium Conference Record,2003IEEE 2468-2472,

[8] J. Cohen (2004). "Bioinformatics – An Introduction for Computer Scientists". ACM Computing Surveys, **36** (2), 122-158.

[9] S. Derksen and H. Keselman (1992). Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables. British Journal of Mathematical and Statistical Psychology, **45** , 265-282.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

AUTHOR'S PROFILE

Jitasha Mishra received her B.Tech in Information Technology From Biju Patnaik University Of Technology in 2008 and M.Tech in Computer Science And Engineering from Biju Patnaik University Of Technology in 2011. She was an Assistant System Manager in Thyssenkrupp Elevator india pvt. Ltd. Between 2008 and 2009, and now is an Assistant Professor in the Department of Computer Science and Engineering at the Gandhi Institute of Technology And Management, Bhubaneswar since 2010. Her research interests are bioinformatics, datamining, hypertext information retrieval, web analysis.

Debashis Hati received his M.tech in Computer Science from NIT, Rourkela in 2000. He was an Assistant Professor in C.V.Raman Engineering college and KIIT University between 2001 and 2010, and now he is an Assistant Professor in the Department of Computer Science and Enineering Gandhi Engineering College. His research interests are bioinformatics, datamining, hypertext information retrieval, web analysis.

Amritesh Kumar received his MCA from IGNOU in 2007 and M.Tech in Computer Science from KIIT University in 2010. He was a trainee in IBM between 2007 and 2008, and he is an Assistant Professor in the department of Computer Science And Engineering at the Gandhi Institute of Technology And Management, Bhubaneswar. Her research interests are bioinformatics, datamining, hypertext information retrieval, web analysis.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Impact of Predicate on Object Oriented Programming

Mohammad Ahmer Munir Khan
Computer science department
ITM University Gurgoan
Gurgoan Haryana, India
ahmermunir2002@gmail.com

Rita Chhikara
Computer science department
ITM University Gurgoan
Gurgoan Haryana, India
ritachhikara@itmindia.edu

A programming language is an artificial language designed to express computations that can be performed by a machine, particularly a computer. Object Oriented programming work around object to relate real world entity on computer. Object is the main active body in real world. Object having a set of attributes and perform some task according to their behavior. The methods in object oriented language are defined to give them behavior of actions. If we provide constraints and logic predicate in term of object and their behavior and attributes, system will start working like actual object in real world and its behaviors can be invoke according to logic and constraints which will provide artificial intelligence in object defined by programmer.

In this paper, we are proposing for a Object Oriented programming language with predicate around that to make programming more realistic and comparative to real world.

Object Oriented Programming , Predicate, Class, Domain declarative programming , imprative programming

I. INTRODUCTION

A programming language is an artificial language designed to express computations that can be performed by a machine, particularly a computer. Programming languages can be used to create programs that control the behavior of a machine, to express algorithms precisely, or as a mode of human communication.

Programming language broadly divided into following category:

- Declarative Programming
- Imperative Programming

Declarative programming is a programming paradigm that expresses the logic of a computation without describing its control flow whereas **Imperative programming** is a programming paradigm that describes computation in terms of statements that change a program state.

Declarative programming is a programming paradigm that expresses the logic of a computation without describing its control flow. Many languages applying this style attempt to minimize or eliminate side effects by describing what the program should accomplish, rather than describing how to go about accomplishing it. Declarative programming often considers programs as theories of a formal logic, and computations as deductions in that logic space.

Declarative programming is an umbrella term that includes a number of better-known programming paradigms.

1. Constraint programming
2. Functional programming
3. Logic programming

Constraint programming

In constraint programming relations between variables are stated in the form of constraints, specifying the properties of a solution to be found. The set of constraints is then solved by giving a value to each variable so that the solution is consistent with the maximum number of constraints.

Constraint programming is often used as a complement to other paradigms: functional, logical or even imperative programming.

Functional programming

Functional languages employ a computational model based on the recursive definition of functions. In essence, a program is considered a function from inputs to outputs, defined in terms of simpler functions through a process of refinement. While functional languages typically do appear to specify "how", a compiler for a purely functional programming language is free to extensively rewrite the operational behavior of a function, so long as the same result is returned for the same inputs.

Logic programming

Logic or constraint-based languages take their inspiration from predicate logic. They model computation as an attempt to find values that satisfy certain specified relationships, using a goal-directed a search through a list of logical rules. Prolog is the best-known logic language. The specifics of how these queries are answered are up to the implementation and its theorem proving.

Imperative programming is a programming paradigm that describes computation in terms of statements that change a program state. In much the same way that imperative mood in natural languages expresses commands to take action; imperative programs define sequences of commands for the computer to perform.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The term is used in opposition to declarative programming, which expresses what the program should accomplish without prescribing how to do it in terms of sequences of actions to be taken

Procedural programming is imperative programming in which the program is built from one or more procedures (also known as subroutines or functions). The terms are often used as synonyms, but the use of procedures has a dramatic effect on how imperative programs appear and how they are constructed. Heavily procedural programming, in which state changes are localized to procedures or restricted to explicit arguments and returns from procedures, is known as structured programming.

Object Oriented Languages is an extension of procedural language which deals with objects. Object-oriented programming (OOP) is a programming paradigm using "objects" – data structures consisting of data fields and methods together with their interactions – to design applications and computer programs. Programming techniques may include features such as data abstraction, encapsulation, messaging, modularity, polymorphism, and inheritance.

II. PREDICATE

A predicate is a verb phrase template that describes a property of objects, or a relationship among objects represented by the variables.

For example, the sentences "The car Tom is driving is blue", "The sky is blue", and "The cover of this book is blue" come from the template "is blue" by placing an appropriate noun/noun phrase in front of it. The phrase "is blue" is a predicate and it describes the property of being blue. Predicates are often given a name. For example any of "is_blue", "Blue" or "B" can be used to represent the predicate "is blue" among others. If we adopt B as the name for the predicate "is_blue", sentences that assert an object is blue can be represented as "B(x)", where x represents an arbitrary object. B(x) reads as "x is blue".

Similarly the sentences "John gives the book to Mary", "Jim gives a loaf of bread to Tom", and "Jane gives a lecture to Mary" are obtained by substituting an appropriate object for variables x, y, and z in the sentence "x gives y to z". The template "... gives ... to ..." is a predicate and it describes a relationship among three objects. This predicate can be represented by Give(x, y, z) or G(x, y, z), for example.

III. IMPACT ON OBJECT ORIENTED PROGRAMMING

Predicate can enhance the scope of object oriented programming language. Some benefits that can be taken out by introducing predicate in object Oriented Language

- Predicate can make Object Oriented Language more powerful.
- Program will be more efficient and corresponding to real world
- Processing of data will be faster
- System can be defined more generic

For Example in a company there are thousands of employee are working. Some of employee having designation manager. We are looking for employees who are manager.

Object Oriented approach:

```
Class Employee {
        boolean isManager;
}
ArrayList<Employee> employees;
Count=0;
for(int i=0;i<employees.size();i++){
        If(employees.get(i).isManager){
                count ++;
        }
}
```

For 1000 employees 1000 clock cycles

Another Approach

```
Class Employee {
        String designation;
}
ArrayList<Employee> employees;
Count=0;
for(int i=0;i<employees.size();i++){
If(employees.get(i).designation.equals("Manager")){
                count ++;
}
}
```

For 1000 employees 1000 clock cycles

Predicate Approach

```
Class Employee{
}
Predicate IsManager(){
        ArrayList managers;
        int getCount(){
                return manager.size();
        }
}
```

Suppose out of 1000 , 20 are manager , then required clock cycle will be 20.

Predicate can help developer to define real world as more generic.We can define the mode of Object and environment in which he/she operating.

For example John is person who can act differently in different conditions.In college john is a faculty in CSE department, here "John gives a lecture to Mary". In home john is care taker of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Tom “John gives a loaf of bread to Tom”. In field John is a boxer “John gives a punch to Jim”
For all scenarios we can define a generic predicate Give(x,y,z).
We can add define different scenarios as clauses and apply predicate on object under different environment.

Environment(College),Faculty(John):Give(John,Lecture,Mary) . Implies John gives lecture to Mary.

Environment(Field),Boxer(John):-Give(John,punch,Jim)
Implies John gives punch to Jim

In Home , John as caretaker ,
Envirnment(Home),Care(John):Give(John,Bread,Tom)

In different environment , john plays different roles , and according to different their action is different, but it can be defined as Give(x,y,z);
Predicate maintains states of Person and perform appropriate task when change in state , but object oriented maintain only data , it cannot maintain state of object. Explicitly we have to pass message to object to initiate the method, by calling give() method on some event.
Introduction of Predicate object oriented language can make object self oriented and perform task according to state of object and environment. The Task can be self initiated according to state and data. Means predicate can make object self intuitive**.**

IV. CONCLUSION

Object oriented languages allows programmer to define entity in term of classes; define behaviors and attributes in term of fields and methods. Behaviors in form of method gives power to programmer to define different activity and fields stores different data related to entity .
But Predicate create more generic way to define different activity of an entity and behaviors of entity under different domain in which entity in working.

REFERENCES

[1] Compilers Principles , Techniques and Tools –Second Edition By Alfred B Aho, Monica S Lam, Ravi Sethi, Jeffrey D Ullman J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Programming Language Pragmatics Second edition –By Michael L. Scott

[3] Modern Compiler implementation in c By Appel

[4] Advanced compiler design & implementation By Muchnick

[5] Compiler Design in C by Allen I. Holub

[6] Compilers: Backend to Frontend and Back to Front Again By Abdulaziz Ghuloum September 17, 2006

[7] http://hubpages.com/hub/Computer-programming-languages-and-generations

[8] http://searchsoa.techtarget.com/definition/object-oriented-programming

[9] http://www.cs.odu.edu/~toida/nerzic/content/logic/pred_logic/intr_to_pred_logic.html

AUTHOR’S PROFILE

Mohammad Ahmer Munir Khan received his Master Degree in Computer Application(MCA) IGNOU in 2006.He is currently pursuing Master of Engineering from ITM University Gurgoan .

Rita Chhikara has Master Degree in Computer Engineering. She is currently pursuing her PhD in Computer science. Currently posted as Assistant Professor in computer engineering department of ITM University.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Framework for Multimedia Data Mining in Information Technology Environment

Owoade A. Akeem
Computer Science Dept. ,
Tai Solarin University of Education
Ijebo Ode, Nigeria.
owoadeakeem@yahoo.com

Ogunyinka T. K.
Computer Science Dept.,
Gateway(ICT) Polytechnic,
Saapade, Remo, Nigeria.
tkogunyinka@yahoo.com

Abimbola B. L.
Computer science Dept.,
Tai Solarin University of Education
Ijebu Ode, Nigeria
bolaikotun@yahoo.com

Abstract: The digital information revolution has brought about profound changes in our society and our lives. The many advantages of digital information have also generated new challenges and new opportunities for innovation which necessitated the mining of multimedia data since multimedia data sets such as audio, speech, text, web, image, video and combinations of several types are becoming increasingly available and are almost unstructured or semi structured data by nature which makes it difficult for human beings to extract information without powerful tools. This call for need to develop data mining techniques which can work for all kinds of multimedia data.

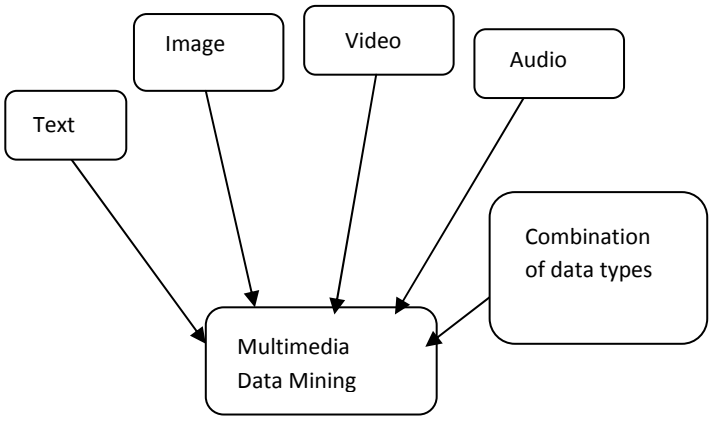
Keywords:

Multimedia data mining, Data mining, Multimedia database.

1. INTRODUCTION

Multimedia data mining is the mining of high level multimedia information and knowledge from large multimedia database. To mine multimedia data we must mine combination of two or more data types such as text and video, or text, video and audio. One solution is to develop mining tools to operate on the multimedia data directly. There has been rapid progress in the field of data mining and data warehousing research, but nothing substantial in mining multimedia. Multimedia has been the major focus for many researchers around the world. Many techniques for representing, storing, indexing and retrieving multimedia data have been proposed. However, rare are the researchers who ventured in the multimedia data mining field. Most of the studies done are confined to the data filtering step of knowledge discovering (KDD) process.

Fig. 1.1 Multimedia data



2.1 Related work:

In A Self Organizing Map (SOM) Extended Model for Information Discovery in a Digital Library Context, Jean-Charles Lamirel et al, 2000 presented a knowledge discovery tool for multimedia databases (or digital

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

libraries), which learning model is an extension of the basic Kohonen self organizing map techniques. The developments are developed and tested on a vesion of an iconographic digital library server developed under the BIBIAN (Bibliographic and Iconographic Based Art Nouveau).

In Data Mining from Functional Brain Images Misturu Kakimoto et al, 2000 studied the discovery of relations between brain areas and brain functions from functional brain images. In addition to the common difficulty of reducing images to a symbolic description, functional brain images challenge the data mining field with the possible existence of correlation between adjacent pixels in an image and the limited number of samples available from a single object. The author applied to real brain images a data mining algorithm developed by Tsukimoto and Morita. The paper presents the results including the discovery of certain rules for finger tapping action and speech related action. In Mining Cinematic Knowledge, Duminda Wijesekera et al, 2006 reported current investigation in an ongoing effort to create a movie mining system. The emphasis of this project is to examine the suitability and applicability of existing data mining concepts to multimedia data, where semantic content is time sensitive and constructed by fusing data obtained from component streams. Smeon J. Simoff et al, 2001 addressed a new area of data mining in collaborative virtual environments. The paper presents a framework for integrating multimedia data mining techniques with collaborative virtual environments, starting from early conceptual development. The aim of presented research was to utilize the multimedia data about the actions and content of collaborative activities in projects conducted in virtual environments, extract meaningful insights out of it and feed discovered knowledge back into environment. The ideas are illustrated with examples from the application of the framework to collaborative workspaces developed in LiveNet, a virtual workspace design system. In the PERSEUS Project: Creating Personalized Multimedia News Portal, (Victor Kulesh et al, 2001) presented the Perseus project. The project is devoted in developing techniques and tools for creating personalized multimedia news portals. The multimedia data mining techniques in this case are used for extracting video clips automatically from TV broadcasts, based on user's preferences. The clips are then augmented with other relevant news from other sources on the Internet. They discussed their approaches to event mining and tracking on the Internet, commercial detection and recognition in video and audio streams, and selection of relevant news video fragments, based on closed captioning and audio transcripts. In Automatic Feature Mining for Personalized Digital Image Retrieval, Kyoung-Mi Lee, 2001 addressed the important issue of measuring similarity based on feature representations of multimedia data. There is an agreement that in its current form the feature space approach does not necessarily represent the notion of similarity in human perception. One of the characteristics of human perception of similarity is that similarity does not vary in the same proportion in all directions in the feature space. Authors presented an incremental method to automatically obtain feature

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

weights based on both the clustered database and on relevance feedback. They presented the results of shape-based indexing and retrieval, showing that using cluster information for an initial search gives better results than using the standard distance. In Relationship Extraction from Large Image Databases, Chabane Djeraba, 2001 addressed the semantic processing of image feature space. The paper presented an algorithm that discovers relationships between image features. Relationships are ranked based on confidence measures. Before the actual mining, the image features (like colors and textures) for a particular database were summarized in a virtual thesaurus. The relationships discovered assisted the automatic categorization of images during their insertion into image databases. At the retrieval stage these relationships were used to improve the accuracy in retrieving relevant images. In Semantic Content-Based Retrieval in a Video Database by Pramod K. Singh, 2001. Authors discussed the issues in managing temporal information of video data that are common to many application areas. The echocardiogram video data management is the specific area addressed in the paper. The paper describes an approach of semantic content-based retrieval of video data using object state transition data model. The advantage of using this model is in allowing storage and indexing of echocardiogram video at different levels of abstraction based on semantic features of video objects. Author presented briefly the system that utilized the proposed approach and discussed the issues in querying the video database. In Data Mining for Typoon Image Collection, Asanobu Kitamoto presented the application of image data mining methods to a narrow domain – the analysis and prediction of typhoons. The image analysis is based on a number of well known techniques, such as principal component analysis, self-organizing maps and time-series analysis to characterize and visualize the statistical properties of typhoon cloud patterns. The prediction is based on the application of an instance-based learning method for analogy-based prediction using past similar patterns. Asanobu emphasized the fundamental problems in typhoon from past similar patterns due to the chaotic nature of the atmosphere. The results tested for this research was the typhoon image collection that was established in the National Institute of Informatics. This medium-size, well controlled and richly-variational collection includes approximately 34,000 typhoon images created from satellite images of a geostationary metrological satellite GMS-5. In Multimedia Data Mining for Traffic Video Sequences, Shu-Ching Chen et al, 2000 presented a framework for multimedia data mining from traffic video sequences recorded at road intersections. Traffic video analysis can discover queues, vehicles identification, traffic flow, and spatio-temporal relations of the vehicles at the intersections including incidents. Several methods are used to analyse the traffic video sequence-background subtraction (a technique to remove nonmoving components from a video sequence), image/video segmentation and object tracking. The spatio-temporal relationships of the vehicle objects in each frame are identified and modeled using multimedia augmented transition networks (labeled directed graphs) are used to derive

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

hierarchical representations of the video clips. The multimedia string represents the transition path in symbolic form that can be processed with some grammar rules. The effectiveness of the proposed methodology was demonstrated with the results of its application on a real life traffic video sequence.

3. MULTIMEDIA DATA MINING ARCHITECTURE

A lot of architectures are being examined to design and develop a multimedia data mining system. The first architecture includes the following: Extract data from the unstructured database, Store the extracted data in a structured database and apply data mining tools on the structured database [8]. This is illustrated in figure 3.1.

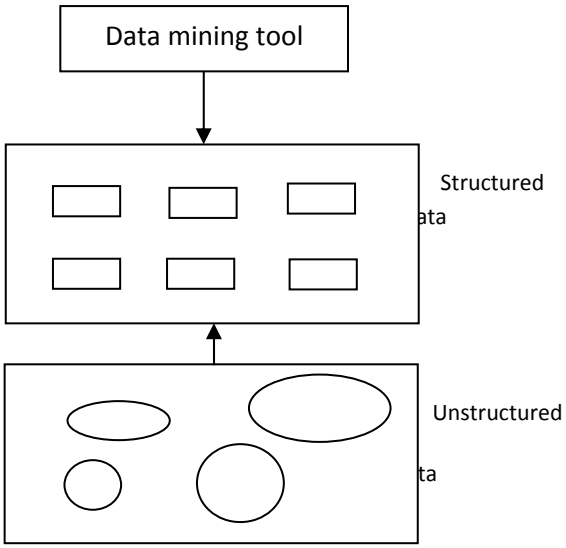


Figure 3.1: Converting unstructured data to structured data for mining

Figure 3.2 present architecture of applying multimedia mining in different multimedia types [1]. Data collection is the starting point of a learning system, as the quality of raw data determines the overall achievable performance. Then, the goal of data pre-processing is to discover important features from raw data. Data pre-processing includes data cleaning, normalization, transformation, feature selection, etc. Learning can be straightforward, if informative features can be identified at pre-processing stage. Detailed procedure depends highly on the nature of raw data and problem's domain. In some cases, prior knowledge can be extremely valuable.

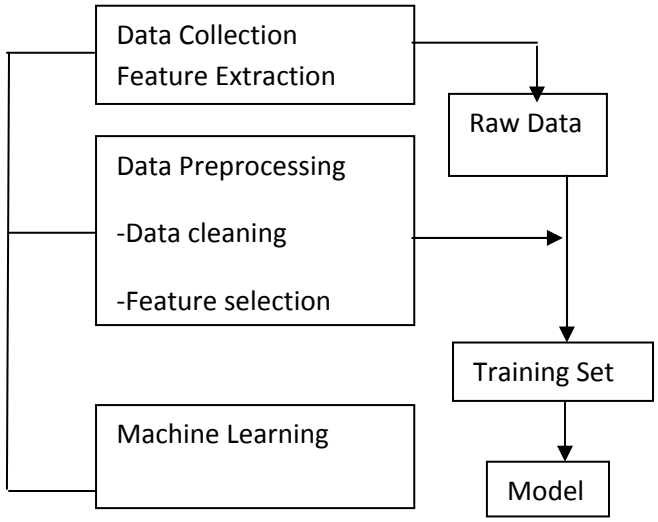


Figure 3.2: Multimedia mining process

For many systems, this stage is still primarily conducted by domain experts. The product of data pre-processing is the training set. Given a training set, a learning model has to be chosen to learn from it. It must be mentioned that the steps of multimedia mining are often iterative. The analyst can also jump back and forth between major tasks in order to improve the results [2]. Figure 3.3 present architecture of applying multimedia mining in different multimedia types [3]. Here the main stages of the data mining process are (4) domain understanding; (5) data selection; (6) leaning and preprocessing; (7) discovering patterns; interpretation; and reporting and using

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

discovered knowledge. The domain understanding stage requires learning how the results of data-mining will be used so as to gather all relevant prior knowledge before mining.

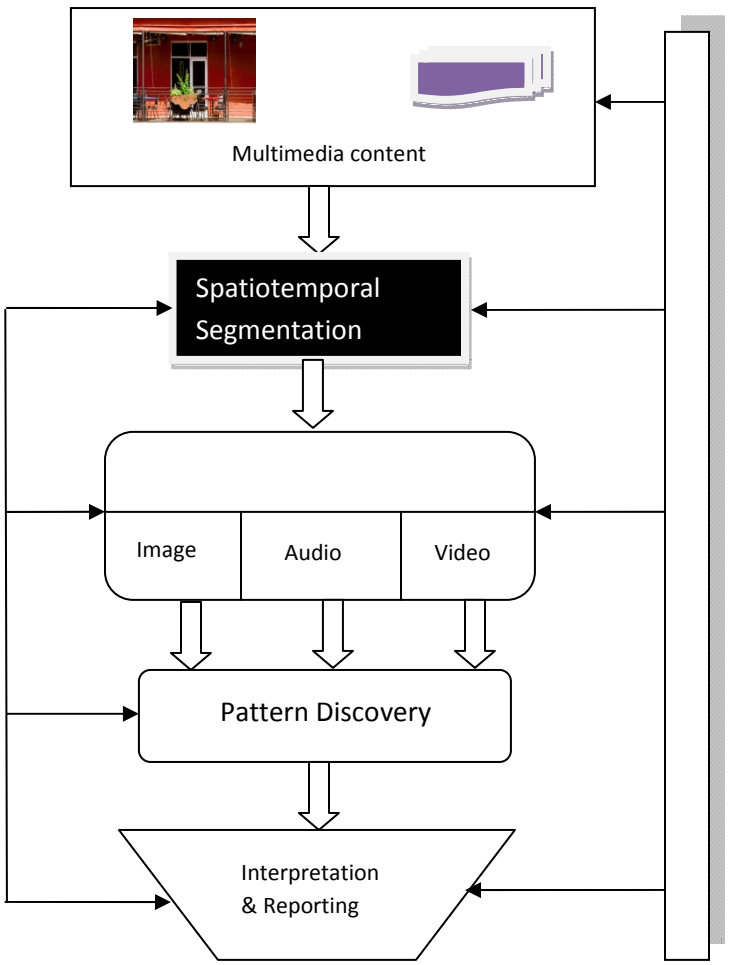


Figure 3.3: Multimedia Data mining architecture

4.0 APPROACHES TO MULTIMEDIA DATA MINING

For multimedia database mining, storage and search techniques need to be integrated with standard data mining methods. Promising approaches include the Construction of multimedia data cubes, the extraction of multiple features from multimedia data, and similarity based pattern searching. *Multimedia data cube* - which facilitates multiple dimensional analyses of multimedia data, primarily based on visual content. A multimedia data mining system prototype, Multimedia Miner has been designed and developed which includes the construction of a multimedia data cube which facilitates multiple dimensional analysis of multimedia data, primarily based on visual content and the mining of multiple kinds of knowledge, including characterization (summarization), discrimination (comparison), classification, association and clustering, in image and video databases. *Feature extraction* - Feature extraction takes the information contained in multimedia data to extract patterns and derive knowledge from large collections of images, audio, video. *Similarity based pattern searching* - Similarity search is a crucial task in multimedia retrieval and data mining. The similarity search is briefly defined as searching for a set of similar objects to a given query object. *Database approach* – the database approach views multimedia data as structured. Features are extracted manually or semi-automatically. The features, referred to as

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

attributes, entail a high level abstraction on unstructured data. the higher the level of abstarction in the features, the lower the scope for ad hoc queries.

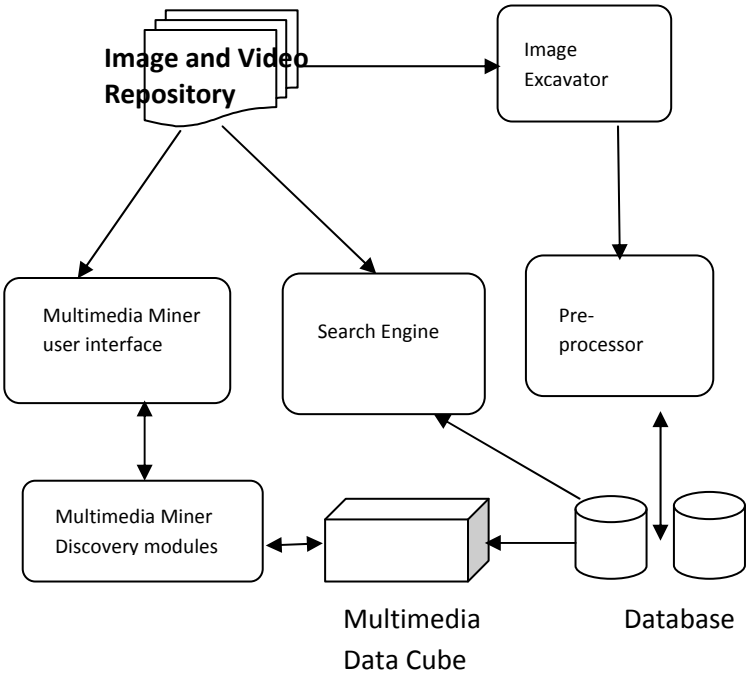


Figure 3.4: General architecture of Multimedia data Miner

Multimedia data mining is the mining of high-level multimedia information and knowledge from large multimedia databases. It includes the construction of multimedia data cubes which facilitate multiple dimensional analysis of multimedia data and the mining of multiple kinds of knowledge, including summarization, classification and association. The common characteristic in many data mining applications, including many multimedia data mining applications is that, first, specific features of the data are captured as feature vectors or tuples in a table or relation and then tuple-mined. There are some examples of multimedia data mining systems. IBM's Query by image content and MIT's Photo book extract image features such as color histograms hues, intensities, shape descriptors, as well as quantities measuring texture. Once these features have been extracted, each image in the database may now be thought of as a point in this multidimensional feature space (one of the coordinates might, for the sake of a simplistic example, correspond to the overall intensity of red pixels, and so on). Another example is MultiMediaMiner. MultiMediaMiner is a system prototype for multimedia data mining which applies multi-dimension database structures, attribute-oriented induction, multi-level association analysis, statistical data analysis, and machine learning approaches for mining different kinds of rules in relational databases and data warehouses. The system contains 4 major components: image excavator for the extraction of images and videos from multimedia

5.0 MULTIMEDIA DATA MINING TECHNIQUES AND ALGORITHMS

The algorithm and techniques employed to perform multimedia data mining are most important. Data mining techniques are numerous. Many of these techniques may also be applied for multimedia data mining. Within the supervised framework, three data mining methods have been used. These are classification, association and statistical modeling. Within the unsupervised learning, clustering is another data mining methodology used.

5.1 Multimedia Data Mining Using Classification Rules: In this approach, we concentrate on discovering the semantic structures. We choose to use the classification rule approaches to perform data mining process because this approach only induce absolutely accurate rules. An early example of this is the work of Yu and

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Wolf, who used one dimensional Hidden-Markov Model for classifying images and video as indoor-outdoor games. A recent work in this area is due to Shu-Ching Chen et al. presented a new multimedia data mining framework for the detection of soccer goal shots by using combined multimodal (audio/visual) features and classification rules using Decision Tree[8].

5.2 Multimedia Data Mining Process Using Clustering: Clustering is a process of organizing objects into groups whose members are similar in some way. It is one of the data mining techniques is an unsupervised learning. In unsupervised classification, the problem is to group a given collection of unlabeled multimedia files into meaningful clusters according to the multimedia content without a priori knowledge. A recent work in this area is due to Lei wang et al, who introduced a clustering method based on unsupervised neural nets and self-organizing maps. Another recent work in this area is due to Jessica Lin et al. have presented an approach to perform incremental clustering at various resolutions, using the Haar wavelet transform using k-means as clustering algorithm.

5.3 Multimedia Data Mining Using Association Rules: Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. There are different types of associations: association between image content and non image content features. An early example of applying association rule mining for image annotation is provided by the work of Ordonez and Omiecinski [9], who consider segmented images to compute the co-occurances of regions that are deemed similar[4]. Another recent work in this area is due to Tseng et al.[10], who proposed a new image classification method by using multiple-level association rules based on the image objects. Another recent work in this area is due to Ankur M. Teredesai et al.[11], who presented a multirelational extension to the FP-tree algorithm to accomplish the association rule mining task effectively. The motivation for using multi-relational association rule mining for multimedia data mining is to exhibit the potential accorded by multiple descriptions for the same image (such as multiple people labeling the sameimage differently).

5.4 Multimedia Data Mining Through Statistical Modeling In this approach, a collection of annoted images is used to build models for joint distribution of probabilities that link image features and key words[4].

CONCLUSIONS

This paper presents multimedia data mining techniques and algorithms. Different approaches to mining multimedia data was discussed with applications.

REFERENCES

1. S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Multimedia Mining, (2004). 'Multimedia mining', WSEAS Transactions on Systems,. Vol. 3, No. 10, pp.3263–3268. Yu H, Wolf, Scenic classification methods for image and video databases. In In SPIE International Conference on Digital Image Storage and Archiving Systems, Vol. 2606,1995,pp. 363-371
2. S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Multimedia Mining, WSEAS Transactions on Systems, Issue 10, Volume 3, December 2004, pp. 3263-3268
3. Valery A. Petrushin and Latifur Khan, Multimedia Data Mining and Knowledge Discovery", 2007 - London: Springer-Verlag, pp. 3- 17.
4. Manjunath et al. "A survey on multimedia data mining and its relevance today", International Journal of Computer Science and Network Security, Vol. 10, No. 11, November 2010.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

5. Bhavani Thuurasingham, Managing and Mining Multimedia, DataBases, Published by CRC Press, 2001.
6. Sanjeevkumar R. Jadhav, and Praveenkumar Kumbargoudar, “Multimedia Data Mining in Digital Libraries: Standards and Features” in Proc. READIT-2007, p. 54.
7. Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, and Jeff Strickrott, "Multimedia Data Mining for Traffic Video Sequences," Proceedings of the Second International Workshop on Multimedia data Mining MDM/KDD'2001), in conjunction with the Seventh ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 78-85, August 26, 2001, San Francisco, CA, USA.
8. Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, Min Chen, A multimodal data mining framework for soccer goal detection based on decision tree logic, International Journal of Computer Applications in Technology**,** Volume 27 , Issue 4 (October 2006), Pages 312-323, year of Publication: 2006.
9. Ordenoz C,Omiecinski E. Discovering association rules based on image content. In:ADL ’99: Proceedings of the the IEEE Forum on Research and Technology Advances in Digital libraries.Washington, DC: IEEE Computer Society; 1999, p.38.
10. Tseng, V.S.; Ming-Hsiang Wang; Ja-Hwung Su, A New Method for Image Classification by Using Multilevel Association Rules, Data Engineering Workshops, 05-08 April 2005 Page(s): 1180 – 1180
11. Ankur M. Teredesai ,Muhammad A. Ahmad, Juveria Kanodia and Roger S Systems”, Volume 10, August, 2006, Springer London.

Owoade Ayoade Akeem is working as an lecturer in the department of Computer Science, Tai Solarin University of Education, Ijebu Ode. He received his first degree in Mathematical Sciences/Computer science from University of Agriculture, Abeokuta, Nigeria in 1998 and master degree in Computer science from University of Ibadan, Nigeria in 2005. He had industrial experience in network monitoring on base station subsystem(BSS) and mobile switching centre(MSC) on ZTE platform at Nigerian mobile Telecommunications limited which took him to ZTE University, China where he attended Advance level course in BSS (2005). He is having total 8 years of industrial and teaching experience. His areas of interests are Multimedia data security, Multimedia data mining and data communications.

Ogunyinka T. K. Obtained his M.Sc in Computer Science at the University of Ibadan, Ibadan, Nigeria. He is a member of both the Nigeria Computer society (NCS) and Computer Profesioner Registration Council of Nigeria (CPN).He is currently a lecturer at the Department of Computer

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Science of Gateway (ICT) Polytechnic,Saapade, Remo, Ogun State. Nigeria.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

DECENTRALIZED INFRASTRUCTURE FOR INFORMATION SHARING USING PEER-TO-PEER TECHNOLOGY.

G. RAMESH KUMAR
Research Scholar, Dept. of Computer Science, Dravidian University, Kuppam, Andhra Pradesh.

Dr. UJWAL A. LANJEWAR
Research Supervisor, HOD, Dept. of Computer Science, Centre Point College, Samarth Nagar, Wardha Road, Nagpur.

Abstract— Peer-to-Peer technology, also known as peer computing, is an emerging paradigm that is now viewed as a potential technology that could provide a decentralized infrastructure for information sharing. The term peer-to-peer refers to the concept that in a network of equals (peers) using appropriate information and communication systems, two or more individuals are able to spontaneously collaborate without necessarily needing central coordination. This paper defines P2P concepts, specifies how P2P is different from Client-Server Model, Distributed Systems and a Grid, and discusses various applications of P2P systems. The main aim of this paper is to review P2P concepts and to highlight its importance through its advantages.

Keywords-component; formatting; style; styling; insert (key words)

I. INTRODUCTION *(HEADING 1)*

Peer-to-Peer technology, also known as peer computing, is an emerging paradigm that is now viewed as a potential technology that could provide a decentralized infrastructure for information sharing. Peer-to-Peer (P2P) has become one of the most widely discussed terms in information technology. The term peer-to-peer refers to the concept that in a network of equals (peers) using appropriate information and communication systems, two or more individuals are able to spontaneously collaborate without necessarily needing central coordination. P2P originally designed exclusively for pragmatic (and legally controversial) file sharing applications, peer-to-peer mechanisms can be used access any kind of distributed resources and may offer new possibilities for internet-based applications.

Many systems have been developed and deployed; e.g., Freenet(6), Gnutella(7), Napster(8), IC!(9), Seti@home(10), LOCKSS(11) and many other. Such architectures are generally characterized by the direct sharing of computer resources (CPU cycles, storage, content) rather than requiring the intermediation of a centralized server. (12)

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

2. DEFINING PEER-TO-PEER CONCEPTS

2.1 PEER-TO-PEER

Peer-to-Peer is a class of applications that takes the advantage of resources storage, cycles, content, human presence available at the edges of the Internet. A peer-to-peer (P2P) architecture is a type of network in which each device has equivalent capabilities and responsibilities. The shared provision of distributed resources and services, decentralization and autonomy are characteristics of P2P networks.

2.2 CLASSIFICATION OF P2P NETWORKS

The classification of the P2P networks according to their degree of centralization

- Pure Peer-to-peer: There is no concept of central sever and central router. Peers act as equals merging both the roles of clients and servers.
- Hybrid Peer – to – Peer: The central server exists to keep information on peers and responds to requests for that information e.g. Gnutella, Free net.

The classification of the P2P networks according to their structure (overlay network links). The P2P network consists of set of peers as network nodes. There are links (directed edge) between the nodes that know the location of each other. Those which have links are structured otherwise unstructured.

- Structured P2P networks : Maintains a distributed hash table and each peer is responsible for a specific part of the content in the network e.g Chord, Pastry, Tapestry, CAN, Tulip.
- Unstructured P2P network : The overlay links are established arbitrarily. There is no correlation between a peer and the content manage by it e.g Gnutella, Fast Track.

2.3 P2P SOFTWARE APPLICATIONS

2.3.1 CHARACTERISTICS OF P2P SOFTWARE APPLICATIONS

The P2P software applications include these seven characteristics:

- The user interface runs outsides of a web browser.
- Computers in the system can act as both clients and servers.
- The software is easy to use and well-integrated.
- The application includes tools to support users wanting to create content or add functionality.
- The application makes connections with other users.
- The application does something new or exciting.
- The software supports : cross-network protocols like SOAP or XML - RPC (19).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

2.3.2 CLASSIFICATION OF P2P APPLICATIONS

P2P architectures have been employed for a variety of different application categories, which include the following

- **Communication and Collaboration:** Systems that provide the infrastructure for facilitating direct, usually real-time, communication and collaboration between peer computers, e.g chat and instant messaging applications. Chat/Irc, Aol, Icq, Jabber.
- **Distributed Computation:** System whose aim is to take advantage of the available peer computer processing power (CPU cycles). e.g Seti@home, genome@home
- **Internet Service, Support:** Systems supporting a variety of Internet services e.g peer-to-peer multicast systems, security applications, virus attacks ect.
- **Database Systems:** Distributed database systems based on peer – to – peer infrastructures e.g LRM, PIER, The Piazza system.

Content Distribution: Systems and infrastructures designed for the sharing of digital media and other data between users. e.g Napster, Gnutella, Freenet.

.

3.COMPARING CLIENT-SERVER, GRIDS, DISTRIBUTEDSYSTEMS WITH P2P

3.1. CLIENT-SERVER AND P2P

In client – Server architecture (fig.1), the client communicates with the server directly and the server maintains client connectivity. By contrast a server does not exist in a peer-to-peer network, because all the devices have the same capabilities and responsibilities (Fig 2), each of them must be able to find other devices and maintain that connectivity within the same network.

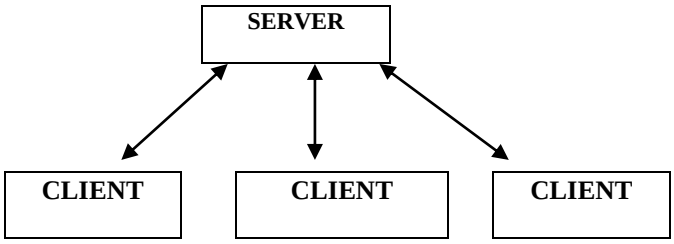


Fig. 1 Client – Server Architecture

Per-to-Peer, is the movement away from the more traditional client-server model to a network where each participating device is acting as both client and server

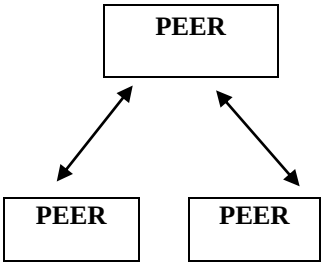


Fig. 2 A Peer-to-Peer Architecture

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

3.2. DISTRIBUTED SYSTEMS AND P2P

The P2P approach is by no means just a technology for file sharing. Rather, it forms a fundamental design principle for Distributed Systems. It clearly reflects the paradigm shift from coordination to cooperation, from centralization to decentralization, and from control to incentives.

Peer-to-Peer systems are distributed systems consisting of interconnected nodes able to self organize into network topologies with the purpose of sharing resources such as content, CPU cycles, storage and bandwidth, capable of adapting to failures and accommodating transient populations of nodes while maintaining acceptable connectivity and performance, without requiring intermediation or support of a global centralized ser authority.

3.3. GRID AND P2P

Grid and P2P both are new emerging approach Distributed computing to address the problem organizing large scale computational. Current Grids provide many service moderate – sized communities and emphasis integration of substantial resources to deliver non qualities of service within an environment of a limited trust e.g. NASA's information power Grid. Contrast current P2P systems deal with many participants e.g Limewire but offer limited specialized services, have been less concerned qualities of service, and have made few if any assume about trust.

4. ADVANTAGES OF P2P

- Unlike the traditional client – server model, where is typically a single or small cluster of server many clients, each node is treated as a peer P2P system and each peer can both consume as provide data and / or services. In addition, each may join and leave the P2P network at any resulting in a truly dynamic and ad-hoc environment.
- P2P applications are categorized into in messaging, file sharing, grid computing collaboration.
- P2P peers provide resources, incl. bandwidth, storage space and computing power.
- P2P networks increase robustness in case of replicating data over multiple users and also enabling peers to find data without relying centralized index server.
- P2P application has the ability to discover, query share content with other peers.
- Peer Reliability : peers in the cluster are distributed and take role accordingly to capabilities.
- P2P is widely used for online multiplayer games.
- P2P can be used for Information sharing and retrieved
- P2P deals with sharing of data files or textual content process data which is continuously created and essential in supporting long-running applications.
- The use of authentication, authorization, encryption establishes trust P2P applications.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- P2P Networking enables or enhances the following scenarios.

Real time communications (RTC): For RTC, peer-to-peer networking enables serverless instant messaging and real-time matchmaking and game playing.

Collaboration: For collaboration, peer-to-peer networking allows the sharing of workspace, files, and experiences.

Content distribution: Peer-to-peer networking allows the distribution of text, audio, and video and software product updates.

Distributed processing: Peer-o-Peer networking allows computing takes to be distributed and processor resources to be aggregated.

Improved internet technologies: Peer – to – Peer networking can also provide an improved utilization of the Internet and support new internet technologies.

- P2P networks are not only used in computer science discipline but also an various disciplines like bioinformatics, education and Academics, Military, Business, telecommunications etc.

5. FUTURE SCOPE

Decreasing costs for the increasing availability of processor cycles, bandwidth, and storage accompanied by the growth of the internet have created new fields of application for P2P networks. In future the concept of P2P will continue to evolve and many P2P applications will be introduced. As peer-to-peer technologies move into more sophisticated and complex applications, such as structured content distribution, desktop collaboration, and network computation, it is expected that there will be a strong convergence between peer-to-peer and Grid computing.

6. CONCLUSION

In this paper, P2P concepts and its comparison with Client Server Model, Distributed Systems, and a Grid have been specified and in the end discuss its importance through its advantages**.**

7. REFERENCES

1. Xiaolin Pang, Barbara catania, Kian-Lee Tan, “Securing You data in Agent-Based P2P systems”.
2. Schoder,D, Fischbach,K, & Teichmann,R (Eds) (2002). Peer-to-Peer Okonomische, technologische und juristische perspektiven.Berlin:Springer.
3. Shirkey,C,truelove,k.,Dornfest,R.,Gonze,l,&Dougherty,D(Eds)(2001).“P2Pnetworking Overview” Sebastopol, CA:O’ Reilly.
4. Schoder, d.,&Fischback,K. (2003). Peer-to-Peer. Prospects. Communications of CM, 46(2), 27-29.
5. Peer-to-Peer, harnessing the power of disruptive technologies, O’ reilly, edited by Andy Oram.
6. Freenet homepage http://freenet. Sourceforge. Com/
7. Gnutlla Development Home Page http://gnutells.wego.com/
8. Napster Homepage http://www.napster.com/
9. ICQ Homepage http://www.icq.com/
10. SETI@home http:// sctiathome.ss. berkeley.edu/

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

11. LOCKASS Home page http:// locakss stanford.edu/
12. Stephanos Androutsellis-Theotokis and Doimidis Spinellis, "A Survey of Peer-to-Peer Content Distribution Technologies, 2004, ACM
13. Shirky,C. What Is P2P….. and What Isn't, 2000
14. Lin Ma, "Develop P2P applications with device discovery technologies", 2005, IBM, retired from www.ibm.org.
15. Miller, M92201) Discovering P2P San Francisco: Sybex
16. Barkai,D. (2001) Peer-to-Peer computing technologies for sharing and collaboration on the net Hillsboro, OR:Intel Press.
17. Abrer, k.,& hauswirth,M (2202). An overview on peer-to-peer information systems. Retrieved from http://Isirpeople.epfl.ch/hauswirth/paper/WDAS2002.pdf.
18. Schollmeier,R,(2002) a definition of Peer-to-Peer networking for the classification of peer-to-peer architectures and applications Proceedings of the first International Conference on Peer-to-Peer computing, 27-29
19. Foster, I, The Grid: a new infrastructure for 21st century science. Physics today, 55(2), 42-47,2002
20. Chien,a, Calder,B, Elbert, Bhatia, K.Rntropia: architecture and performance of an enterprise Desktop grid System, Journal of parallel and distributed Computing.
21. Oram, A. (ed), peer-to-peer: harnessing the power of disruptive technologies. O' Reilly, 2001.
22. Czajkowski,K., Fitzgerald, S., Forste, I. and Kesselman,C.,Gird Information services or Distributed Resource Sharing. 10th IEEE international Syposium on high Performance Distributed Computing, 2001, IEEE Press, 181-184
23. Foster, I, kesselman,c.,tsudik,G and Tuecke,S.A security Architecture for Computational Grids ACM Conference on computers and Security, 1998,83-91.
24. Lai,C., Medivinsky,g and Neuman, B.C. Endorsements,Licensing,and Insurance for Distributed System services, Proc 2nd ACM Conference on computer and Communications Security, 1994.
25. Ian Foster, Adriana Iamnitchi," On Death, taxes, and the Convergence of peer-to-peer and Grid computing".

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Performance Evaluation for Scalable Recursive Multicast Protocol using ns2 simulator

[1]Jafar Ababneh,[2]Firas E.Albalas,[1]Nidhal Kamel Taha El-Omari,[1]Abdel Rahman A.Alkarabsheh, [1]Abd Alsalam Obiadat, [3]Mahmood Baklizi

[1]Faculty of science and information technology, The World Islamic Sciences and Education (W.I.S.E.) University, Amman, 11947, P.O. Box 1101, Jordan
[1]{ jafar.ababneh, nidhal.omari, Ar.karabsheh}@wise.edu.jo
[2]Faculty of science and information technology, Jadara University, Amman – Irbid main Road, 21110 , P.O. **Box 733, Jordan**
[2]**falbalas@jadara.edu.jo**
[3]National advanced IPV6 center (NAV6) university sains Malaysia,11800 USM, penang, malaysia
[3]mbaklizi@wise.edu.jo, mbaklizi@nav6.org

ABSTRACT

In multicast routing the scalability issue should be considered, this issue comes because the increasing in the size of the Multicast Forwarding Table (MFT) because of the increase in multicast group members or the increase in the number of multicast groups. SReM[1] is a multicast routing protocol that addressed this issue by explicitly encode the building the multicast tree.

An extensive evaluation performance is considered for this protocol in this paper. As a result, this protocol gave an improvement in the scalability issue by minimizing the header size and gave an improvement in the packet delivery ration and the end to end delay.

INTRODUCTION

Due to the scarce of resources in the internet, multicast provides an efficient solution to use these recourses fully and efficiently. Also because of the explosive increasing in traffic over the Internet, multicasting became an important issue in routing protocols [2, 3]. Some issues or the increase in the number of multicast groups. So there are two main aspects that can be used to evaluate the scalability in multicast protocols: scalability with regard to the number of group members and to the number of multicast groups in the network. Recently, there are a number of proposed mechanisms to solve the scalability issue in multicast protocols, which can be categorized into tunnelling techniques, forwarding state reduction and explicit multicast.

The main aim for this paper is to perform an extensive performance evaluation of a novel explicit multicast protocol; Scalable Recursive Multicast Protocol (SReM) [8];and compare this protocol with most well-known explicit multicast protocols.

SREM Overview

Due to the scarce resources in the Internet, multicast provides an efficient solution to use these recourses fully and efficiently. Because of the explosive increasing in traffic over the Internet, multicasting became an important issue in routing protocols [2, 3]. Some issues are still open such as scalability [4], billing [5], address allocation [6] and security [7], where the scalability has drawn much attention for Internet application.
In this paper a protocol called SReM[1] is discussed and an extensive performance

Corresponding author. E-mail address: jafar.ababneh@wise.edu.jo

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

evaluation using ns2 simulator [9]is considered. The basic idea behind SReM is to forward data between dynamically selected nodes called Branching Node Routers (BNRs). Its goal is to reduce the state information kept for the multicast group members and to reduce the routing overhead by providing a local join/leave and tree maintenance procedures using fixed size messages. Hence, it is achieving higher degree of scalability. A detailed description and the detailed SReM process in joining and leaving nodes can be found in [1]and a detailed cost analysis for this protocol can be found in [1].

PERFORMANCEEVALUATION

Performance evolution is an idea used to measure and evaluate the performance of routing protocols. The simulation runs on different scenarios and evaluating different metrics. The results obtained from this part of evaluation are also compared with other protocols in the same area of SReM.

SIMULATION ENVIROMENT

The proposed protocol is implemented using ns2 simulator [9] (version 2.29). The simulation environment consists of 60 nodes, these nodes represent the number of LMRs, and each LMR can carry any number of receivers depending on its configuration. In our simulation, each LMR can connect up to 10 receivers so the maximum number of receivers is 600 nodes. The number of multicast group members (LMRs) varies from 5, 10, 15, 20, 25, 30, 35, 40, and 45 nodes; these nodes represent LMRs in SReM and Designated Router (DR) for xcast+ protocol.

Traffic generation considered at this simulation is CBR traffic with payload size 512 bytes. Data packets are generated at source at a rate of 8 packets per second; this will introduce 4096 byte per second. Each simulation runs for 200 second.

In this protocol evaluation, SReM is compared with Xcast+ protocol, the selection of Xcast+ is done for the following reasons:-

- Xcast+ is widely referenced in the area of explicit multicast protocols.
- This protocol (Xcast+) is relatively close to the proposed protocol (SReM), so the comparison will give good indication of the performance of SReM.

PERFORMANCE METRICS

The following metrics are used to evaluate the performance of proposed work:-

- Average Packet Header Size represents the size of the header included in the data packet in bytes. It represents the size of each data packet header in order to deliver this packet to all destinations.
- Average End To End Delay: is the average time that takes the data packet sent by source node to reach its destination node.

Total end to end delay = SUM (time_received(pkt) – time_sent(pkt)) for all data packets

Avg. End To End Delay (AED) = Total end to end delay / no. of data packets received

SIMULATION RESULTS

The simulation results will be discussed in this section, this discussion is organized with regard to the metrics used for protocol evaluation. At each part, the results are presented in figure form and then a discussion and explanation for these figures is mentioned.

Packet header size

Figure 1 shows the results obtained for the first part of packet header size evaluation. It can be noticed that SReM has got static header size even when the group size is increasing. This result comes because whatever the group size SReM will only include the next Branching Node Router (BNR) addresses in the header of each data packet. In Xcast+ the results show that an exponential increase of header size when the group size increases. In conclusion, SReM improve the scalability feature because the

Corresponding author. E-mail address: jafar.ababneh@wise.edu.jo

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

header size is constant even when the group size increases.

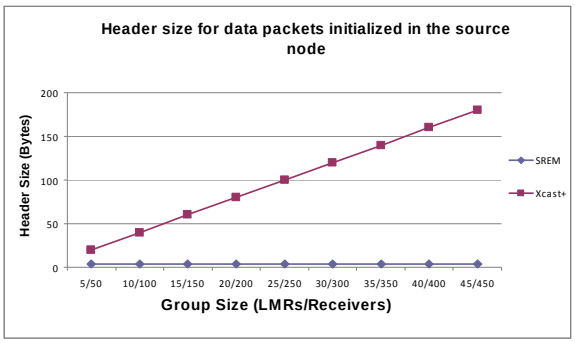


Figure (1) Extra packet header size as a function of group size

The results for the second part of evaluation is shown in Figure 2, the size of packet header for SReM increases slightly but in Xcast+ a high increase of packet header is happens again. This result proofs that SReM improves the scalability feature in wired networks.

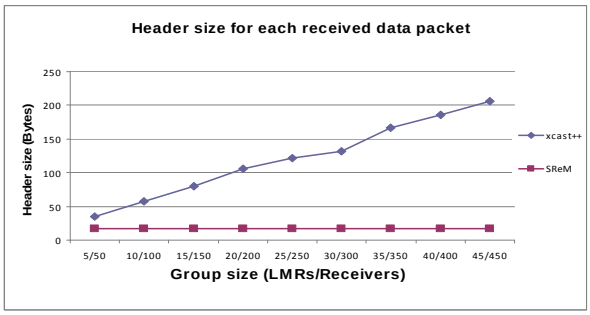
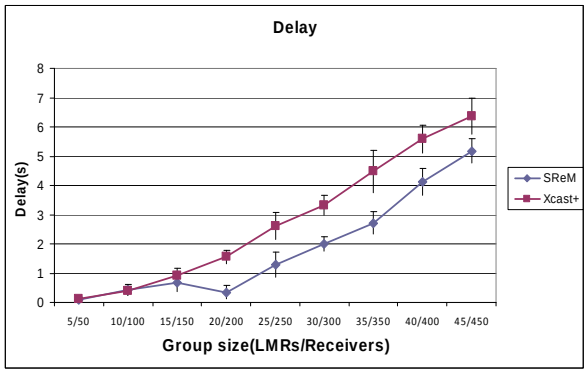


Figure (2) Average header size as function of group size

End To End Delay

Figure 3 shows the average end to end delays as a function of group size with error rate because of random generation for simulator. The bigger value of delay the less efficient the protocol. In reality, SReM shows lower delay in comparison with Xcast+. This is because scalability feature in SReM make the intermediate nodes to forward the data packets faster where in Xcast+ the data packets will take high processing time in the intermediate nodes which lastly will increase the delay for the data packets arrival at the end nodes.



Figure(3) End to End Delay.

CONCLUSION

In this paper a performance analysis study is done for SReM[1]. SReM is a scalable multicast protocol for fixed networks. The scalability issue is an important feature for wireless networks.

The performance analysis shows that SReM scales well when the multicast group size becomes large. The results show that SReM performs a fixed header size in data packets where the header size for Xcast+ protocol increases exponentially when the group size increases. These results show that SReM improves the scalability feature in networks. Other results obtained shows that SReM performs less end to end delay and overhead in addition to scalability feature.

REFERENCES

1. Cao, Y. and K. Al-Begain. *SREM: A Novel Multicast Routing Algorithm-Comprehensive Cost Analysis*. in*5th World Wireless Congress (WWC 04)*. 2004. San Francisco, USA.
2. Ballardie, T., P. Francis, and J. Crowcroft, *Core Based Trees (CBT) An Architecture for Scalable Inter-Domain Multicast Routing.* Computer Communication review, 1993. **23**: p. 85-85.
3. Waitzman, D., C. Partridge, and S.E. Deering, *Distance Vector Multicast Routing Protocol.*RFC Editor 1075 , United States, 1988.
4. El-Marakby, R. and D. Hutchison. *Scalability Improvement of the Real-*

Corresponding author. E-mail address: jafar.ababneh@wise.edu.jo

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

time Control Protocol (RTCP) Leading to Management Facilities in the Internet. in*Computer Communications*. 2005.
5. Dondeti, L., et al. *MBA: a tool for multicast billing and accounting*. in*Proceedings Sixth IEEE Symposium on Computers and Communications*. 2001.
6. Bhattacharyya, S., D. Towsley, and J. Kurose. *The loss path multiplicity problem in multicast congestion control*.in*18th Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM'99*. 1999.
7. Judge, P. and M. Ammar, *Security issues and solutions in multicast content distribution: a survey.* IEEE Network, 2003. **17**(1): p. 30-36.
8. Al-Begain, K., Y. Cao, and K. Alameh, *A DBT-Based Mobile Multicast Protocol.* Systems Communications., 2005: p. 255-260.
9. *The network simulator - NS2* in *URL:*http://www.isi.edu/nsnam/2005.

Authors Information

Dr. jafar Ababneh is an assistant professor. He received his PhD degree from Arab Academy for Banking & Financial Sciences (Jordan) in 2009. He received his M.Sc degree in computer engineering from University of the Yarmouk (Jordan) in 2005. He earned his B.Sc in Telecommunication engineering from University of Mu'ta (Jordan) in 1991. In 2009, he joined The World Islamic Sciences and Education (WISE) University as a head of the departments of computer information systems and network systems in the school of information technology(CISN). He has published many research papers and book chapters in different fields of science in refereed journal and international conference proceedings.

His field research lies in development and performance evaluation of multi-queue nodes queuing systems for congestion avoidance at network routers using discrete and continuous time, also his research interests includes computer networks design and architecture, wire and wireless communication, artificial intelligence and expert system, knowledge base systems, security systems, data mining and information.

Corresponding author. E-mail address: jafar.ababneh@wise.edu.jo

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Mrs Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Mr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Mr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Mr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Mr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Mr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Mr. P. Vasant, University Technology Petronas, Malaysia
Mr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Mr. Praveen Ranjan Srivastava, BITS PILANI, India
Mr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Mr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Mr. Tirthankar Gayen, IIT Kharagpur, India
Ms. Huei-Ru Tseng, National Chiao Tung University, Taiwan

CALL FOR PAPERS
International Journal of Computer Science and Information Security
January - December
IJCSIS 2012
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500